LISBOA
SCHOOL OF
ECONOMICS &
MANAGEMENT

Master in
Actuarial Science

Masters Final Work
Dissertation

Analysis of the Claims Data of a Life Insurance Portfolio

Ana Luísa Mendes Lima Pereira

May - 2016

LISBOA
SCHOOL OF
ECONOMICS &
MANAGEMENT

Master in
Actuarial Science

Masters Final Work
Dissertation

Analysis of the Claims Data of a Life Insurance Portfolio

Ana Luísa Mendes Lima Pereira

Supervisors:
maria da conceição esperança amado
Onofre Alves Simões

May - 2016

Acknowledgements

Firstly, I would like to thank my advisers Professors Conceição Amado and Onofre Simões. Over a year ago, I had the idea of bringing together what I had learned in the masters with a field I knew very little of but was very interested in (data mining) and they both played a crucial part in making that come true. Thank you for your time, patience and all your feedback throughout this process.

Secondly, I would like to thank my employer and specifically my bosses for making the masters possible and for their support in my continued education. I would like to thank my colleges, both my past team (Marco and Ana Patrícia) who played an important part in the beginning of this work and my current team (Filipa, Vasco, Hugo and Marisa) for their support and motivational words. This thesis would not be the same without all that I've learned these past four years.

Last but not least, I have to thank all my endlessly supportive boyfriend Mário, my wonderful friends and my family who helped me pull through these past three years of insanity of working and studying. You have my deepest gratitude.

In the words of Bob Dylan, "And that's really all. If I've had anything to tell anybody, it's that: You can do the impossible. Anything is possible. And that's it. No more.".

Abstract

Mortality graduation is a problem that has long been studied in actuarial science using many different approaches. Creating a graduation specific to a portfolio is common practice amongst insurance companies since it can be a powerful tool for estimating future mortality, especially in a large portfolio, as the one available for this study.

In this thesis, we will focus on a number of specific examples of techniques for graduation: Gompertz's law; an empirical approach where mortality rates are fit to an exponential curve; graduation by standard table, using the Swiss tables GKM/F80 and GKM/F95; generalised linear models (GLM); three different types of regression trees - classification and regression trees (CART), conditional inference trees and random forests. Furthermore, following Guo *et al.* (2002), hybrid methods will be created, some of which unseen before in the literature, combining regression trees with some of the other approaches.

These techniques will be applied to the traditional concept of mortality rate but also to a version of it weighted by sum assured. This is done by attaching weights (the sum assured) to each death and unit of exposure to risk. As expected and previously observed in literature, mortality will generally be lighter for policies with higher sum assured, which is in line with the idea that people with higher sums assured are wealthier and hence healthier, living longer lives.

Whereas mortality studies often encompass only the age and gender of the insured lives as explanatory variables, for this study, the civil status and place of residence were also available. These explanatory variables proved relevant when applying the tree generating algorithms. In the end, we will find that both for the traditional and the weighted mortality rates, a hybrid method (of a regression tree with the empirical approach applied to its leafs) yielded the best results for the portfolio in study. The RMSE (root mean square error) was the evaluation metric used.

Keywords: Graduation, Mortality Rates, Mortality Rates Weighted by Sum Assured, Regression Trees

Resumo

A graduação de mortalidade é um problema há muito estudado nas ciências atuariais, utilizando muitas abordagens diferentes. Criar uma graduação específica a um portfolio é uma prática comum na indústria, dado que pode ser uma ferramenta poderosa para estimar a mortalidade futura, especialmente em carteiras de grande dimensão, como foi o caso desta.

Para este estudo, vamos focar-nos em alguns exemplos específicos de graduação: a lei de Gompertz's; uma abordagem empírica em que as taxas de mortalidade são ajustadas a uma curva exponencial; graduação por tábua de mercado, utilizando as tábuas Suiças GKM/F80 e GKM/F95; modelos lineares generalizados (GLM); três tipos diferentes de árvores de regressão - árvores de classificação e regressão (CART), árvores de inferência condicional (*conditional inference trees*) e florestas aleatórias (*random forests*). Além disso, seguindo Guo *et al.* (2002), serão propostos métodos híbridos, alguns dos quais ainda não existiam na literatura, combinando árvores de regressão e as outras técnicas.

Todos estes modelos serão aplicados tanto ao conceito tradicional de taxa de mortalidade como a uma sua versão ponderada por capital seguro. Tal é feito ao multiplicar cada unidade de morte e exposição ao risco pelo respetivo capital seguro. De acordo com o esperado e a literatura, a mortalidade será mais baixa para apólices com capital seguro maior, o que vai de encontro à ideia de que pessoas com maior capital seguro estarão numa situação económica mais favorável e portanto conseguirão viver mais tempo.

Enquanto a prática comum em estudos de mortalidade consiste em utilizar apenas a idade e o género como variáveis explicativas, para este estudo, o estado civil e local de residência também estavam disponíveis. Estas variáveis explicativas revelaram-se importantes quando aplicados modelos de árvores de regressão. Conclui-se que para qualquer uma das taxas de mortalidade estudadas o melhor modelo correspondeu a um modelo híbrido que combina uma árvore de regressão com a abordagem empírica aplicada às suas folhas. Como métrica de comparação foi utilizado o RMSE (raiz quadrada do erro quadrático médio).

Palavras-Chave: Graduação, Taxas de Mortalidade, Taxas de Mortalidade Ponderadas por Capital Seguro, Árvores de Regressão

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem description

For this thesis, we will focus on modelling mortality data, with the purpose of creating an effective graduation. By graduation we refer to the set of principles and methods by which the observed (or crude) probabilities of death are fitted to provide a smooth basis for making practical inferences and calculations of premiums and reserves, as defined in Debón *et al.* (2005).

After calculating the crude mortality probabilities, graduation is necessary in order for the final probabilities to be plausible, since the observed values usually present brusque changes between consecutive ages or drop to zero when no deaths were observed. We will focus on two types of graduation methods:

- Graduation by parametric formula: models for which the data is adjusted to a function, making assumptions about the distribution of the data;
- Graduation by non-parametric formula: models for which the mortality probability does not take a predetermined form but is constructed according to information derived from the data.

The general methodology is essentially the same for the two methods, namely calculating the crude probabilities, choosing a model, fitting it and testing the graduation. Each method can produce many possible graduations and the best one should be chosen according to a measure of adherence to reality and smoothness.

The main purpose of this work is to use examples of each type of graduation as well as combinations of these and evaluate the models according to adherence. Furthermore, whereas traditionally graduation is done focusing on the personal characteristics age and gender alone, we will introduce also the insured lives' civil status and place of residence. Mortality will also be studied from two different perspectives: mortality rates calculated using the number of deaths and the exposed to risk (the traditional approach); and mortality rates weighted by sum assured, calculated using the sum assured of the recorded deaths and the risk exposure weighted by sum assured.

1.2 Literature review

One simple model for graduation by parametric formula is Gompertz's law, from Gompertz (1825), which assumes that the force of mortality follows an exponential curve, growing with age. This constitutes one of the most influential proposals from the early times of mortality modelling and is still very useful, as stated in Bowers *et al.* (1997). Many contributions in the field of mortality laws generalise or proceed from Gompertz *et al*'s ideas. Remarkable examples are given in Makeham (1860) (adds an age independent component to the exponential growth), Thiele (1871) (a seven parameter formula which covers the whole span of life) and Oppermann (1872) (a three parameter model for ages bellow or equal to 20).

As explained in Haycocks and Perks (1955), another class of parametric models is graduation by reference to a standard table, for which a standard table is adjusted to the observed data, by way of a simple transformation such as a percentage or a shift in age.

McCullagh and Nelder (1989) introduced another type of parametric model called Generalised Linear Models (GLM). The use of GLM for the graduation of both the probability of death and the force of mortality is justified because both response variables are not normal. The experience in graduation using GLM has been compiled in actuarial literature by Renshaw (1991) and Verrall (1996).

Regarding non-parametric graduation, one such example is kernel estimation. In Gavin *et al.* 1993, a link between moving weighted averaged graduation and kernel estimation is explored and new kernel estimator for graduation is studied and an optimal smoothing kernel derived. We will, however, be focusing on another example of non-parametric graduation: regression trees. Tan *et al.* 2006 offers a general introduction to the concepts behind this technique whereas Kim *et al.* (2011) and Chapados (2010) show possible applications in the context of insurance. The former aimed to assess whether the performance of various data mining techniques, such as the artificial neural networks, support vector machines and decision trees, outperform logistic regression for mortality prediction in an intensive care unit. The latter details the technology behind statistical learning algorithms and data mining (namely artificial neural networks) for estimating the pure premium of an insurance contract and then applies these techniques to a real-world automobile insurance pricing project.

In this thesis, we propose the use of a new approach based on hybrid models of several of the above methods, inspired by Guo *et al.* (2002), a paper which addresses issues and techniques for the study of advanced age mortality. More specifically, in this paper the influences of the available explanatory variables on mortality were identified with exploratory data analysis and decision tree algorithm. Afterwards, models to address their effects were built with logistic regression.

Regarding the modelling of mortality rates weighted by sum assured, examples can be found in Roberts (1993, 1992). These show how calculating the ratio between weighted

mortality rates and traditional mortality rates provides a straightforward way for an insurance company to monitor the underlying mortality of its portfolio over time and acts as an early warning sign for possible deterioration of underwriting results.

Both approaches to the study of mortality (weighted and traditional rates) are used in industry standards such as the Continuous Mortality Investigation, as stated in SAPS Mortality Committee (2008, 2009).

1.3 Thesis outline

This thesis is organised in six chapters, as follows. Chapter 1 introduces the motivation of the work, with the task of graduating mortality probabilities being introduced and briefly discussed. Chapter 2 describes the concepts implicit to mortality graduation and each of the models that will be applied. In Chapter 3, the portfolio used throughout this study is described and transformed in order to finally allow for the analysis of its mortality rates. Chapter 4 presents the results obtained with the various models described in Chapter 2 and a comparison of the results between them. Chapter 5 has examples of an actuarial application using the best graduations in Chapter 4. Finally, Chapter 6 summarises the main findings of this thesis, discussing the advantages and disadvantages of each tested model. Further developments to the work are also discussed.

Chapter 2

Mortality Graduation

2.1 Basic concepts in mortality theory

In this section, the concepts behind the theory of mortality will be explained. The notions exposed bellow are well established and can be found in Bowers *et al.* (1997).

Consider T, the positive random variable representing the complete life duration of an individual. For that individual, $\{T_x : x = 0, 1, 2, ..., w\}$ is defined as the residual life time given that the individual reaches age x (consider w as the limiting age, which no individual will pass). Then,

$$P(T_x > t) = P(T_0 > x + t | T_0 > x),\ t > 0.$$

Using this concept, define the instantaneous mortality rate (or force of mortality) at age $x + t$ as

$$\mu_{x+t} = lim_{\Delta t \to 0^+} \frac{P(t < T_x \leq t + \Delta t | T_x > t)}{\Delta t}.$$

The probability of an individual aged x dying before or on age $x + t$ (for $t \geq 0$) is defined as

$${}_tq_x = P(T_x \leq t) = P(T_0 \leq x + t | T_0 > x).$$

The above is called probability of death and when written as q_x, implies $t = 1$, i.e., it's the probability of an individual aged x dying before age $x + 1$, in which case it is also called the mortality rate at age x. It can be established (see Bowers *et al.* (1997)) that q_x is related to the force of mortality through the equation

$${}_tq_x = 1 - exp\left(-\int_0^t \mu_{x+y} dy\right).$$

As the data for mortality if usually available only for $x = 0, 1, 2, ..., w$ and it is often necessary to calculate the probability of death for ages or intervals of time which are not integers, for instance ${}_{1-u}q_{x+u}$ with x an integer and $0 \leq u \leq 1$, assumptions must be made as to how the death data behaves between discrete points in time. For this thesis, we will

be using the so called constant force of mortality hypothesis. This hypothesis assumes that between ages the force of mortality is constant, i.e., $\mu_{x+u} = \mu_x$, for $0 \leq u < 1$, and the probability of death is therefore given by

$$ {}_uq_x = 1 - exp\left(-\mu_x \times u\right),\ 0 \leq u < 1. $$

Consider now an individual who was insured from age $x+t$ to age $x+s$, where $0 \leq t \leq s$. The central exposure to risk of this individual at age x, E_x, is defined as the observed total time at risk of the individual at that age. Given that the individual was insured from age $x+t$ to age $x+s$, then $E_x = s - t$. Extending the concept to a group of n individuals, each insured from age $x+t_i$ to age $x+s_i$ $(i = 1, ..., n)$, the central exposure to risk for all individuals aged x is given by

$$ E_x = \sum_{i=1}^{n} \left(s_i - t_i\right). \tag{2.1.1} $$

Next, define D_x as the random variable representing the number of deaths for individuals aged x and d_x as its value for a given x. One common model for this random variable, under the constant force of mortality hypothesis, is the Poisson distribution with parameter $\mu_x \times E_x$. Using this model, the maximum likelihood estimate of μ_x is $\hat{\mu}_x = \frac{d_x}{E_x}$.

We will be working to model

$$ \hat{\mu}_x = \frac{d_x}{E_x},\ q_x = 1 - exp\left(-\hat{\mu}_x\right) = 1 - exp\left(-\frac{d_x}{E_x}\right). \tag{2.1.2} $$

Throughout the rest of this thesis, whenever "mortality rates" are referenced, without explicitly saying that they are weighted, we will mean the traditional version, as described in this section.

2.1.1 Weighted mortality rates

According to Roberts (1993), in some particular cases it is sometimes useful to consider the use of weighted mortality rates. This is done by attaching weights to each death and unit of exposure to risk: in life insurance these weights are typically sums assured or numbers of policies. For this thesis, we will be focusing on using the sums assured as the weight.

Weighting the rates in this way is a natural thing to do, in that what matters ultimately to an insurance company is the monetary amounts requiring to be paid out. Multiplying total sums at risk in an age interval, for example, by the central weighted mortality rates yields an estimate of total payments to that group over the next year, provided that the weights remain unaltered.

An early author treating the behaviour of weighted rates such as those considered here is Cody (1941), who finds expressions for the mean and variance of initial weighted mortality rates, and derives the ratio of the variances of the weighted and traditional

rates. More recently, Klugman (1981) has compared the mean square errors of weighted and traditional initial mortality rates, setting off the larger variance of the weighted rates against the bias implicit in using the traditional rates. The paper in question also pointed out that mortality is generally lighter for policies with higher sums insured.

Adapting the terminology from the previous section, define the central weighted risk exposure for an individual aged x as

$$E_x^{SA} = E_x \times SA,$$

where SA is the sum assured for that individual at age x. Given that an individual can have more than one policy (hence, sum assured) and we will assume an evolution of the sum assured at the start of each new civil year, a more complete definition is given by

$$E_x^{SA} = \sum_{policy=1}^{n} \sum_{year=1}^{m} E_x^{policy,year} \times SA^{policy,year},$$

where n is the number of policies of the person in question and m is the number of civil years during which the individual was insured. If the sum assured for an insured live is 1 unit per annum throughout a period and there is only one policy, the unweighted and weighted exposures will be equal.

Extending the concept to a group of k individuals, each with n_i policies $(i = 1, ..., k)$ over m civil years, the weighted central exposure to risk for all individuals aged x is given by

$$E_x^{SA} = \sum_{i=1}^{k} \sum_{policy=1}^{n_i} \sum_{year=1}^{m} E_x^{policy,year} \times SA^{policy,year}. \tag{2.1.3}$$

Conversely, instead of using d_x, the number of deaths at age x, we will be using d_x^{SA}, the total sum assured of the lives who died at age x. If all deaths had a sum assured of 1 unit, then $d_x = d_x^{SA}$.

We will be modelling

$$\hat{\mu}_x^{SA} = \frac{d_x^{SA}}{E_x^{SA}},\ q_x^{SA} = 1 - exp\left(-\hat{\mu}_x^{SA}\right) = 1 - exp\left(-\frac{d_x^{SA}}{E_x^{SA}}\right). \tag{2.1.4}$$

2.2 Models

In this section, a brief description of the models used for mortality graduation is given. Regression analysis is a statistical process for estimating the relationships between variables. The objective is to find a model that uses explanatory variables as input and outputs values of the response variable, the variable being modelled.

2.2.1 Gompertz's law

Gompertz's law, as defined in Gompertz (1825), is as follows: given α and β positive parameters and an age x, the force of mortality for that age can be modelled as $\mu_x = \alpha \times exp(\beta x)$. The mortality rate will then be given by

$$q_x = 1 - exp\left(-\alpha \times exp(\beta x)\right). \tag{2.2.1}$$

This simple law has proved to be a remarkably good model in different populations and in different epochs, and many subsequent laws are modifications of it.

2.2.2 Empirical approach

Another parametric model, similar to Gompertz's law, is modelling the mortality rates, instead of the force of mortality, as an exponential curve. Although this model is not supported by literature, it is commonly used in the insurance industry and will be applied in this thesis.

For this model, given α and β positive parameters and an age x, the mortality rate for that age can be expressed as

$$q_x = \alpha \times exp(\beta x). \tag{2.2.2}$$

2.2.3 Standard tables

Graduation by reference to a standard table can be quite useful when there is not a large amount of data and there is a standard table which seems appropriate. Provided a simple function is chosen and the standard table is smooth to begin with, the resultant rates are automatically smooth.

By standard table we are referring to a published life table based upon sufficient data to be regarded as reliable. In our case, the tables are (see Appendix A):

- GKM80 and GKF80- constructed based on the experience of around 1.900.000 men and 260.000 women during the years of 1971 to 1975;
- GKM95 and GKF95- constructed based on the experience of around 3.800.000 men and 1.540.000 women between 1986 and 1990.

The third letter of each name (M/F) stands for male and female tables.

These Swiss tables were chosen because they are commonly used for life insurance tariffs in Portugal. The adjustment was given by:

$$q_x = \alpha \times GK_x, \tag{2.2.3}$$

where GK_x is the value of the table in question at age x and α is a positive parameter.

2.2.4 Generalised linear models (GLM)

Generalised linear models (GLM) are a generalisation of linear regression models that can be used in certain cases where they are not appropriate. In GLM, made popular in McCullagh and Nelder (1989), there are three components:

- **Random Component**: regards the probability distribution of the response variable Y. The distribution of Y belongs to the exponential family of distributions (binomial, Poisson, normal, *etc*).
- **Systematic Component**: related to the explanatory variables $\mathbf{X} = (X_1, ..., X_k)$ and its linear relation with the predictor $\eta = \beta_0 + \mathbf{X}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}= (\beta_1, ..., \beta_k)$.
- **Link Function**, η or $g(\mu)$: a function which specifies the link between the random and the systematic components. It expresses how the expected value of the response relates to the linear predictor of explanatory variables, i.e., $E[Y] = \mu = g^{-1}(\eta)$.

2.2.5 Regression trees

When trying to find a parametric model for a dataset which has many different variables interacting in complicated and nonlinear ways, finding the right parametrisation can be a difficult task. The option of supervised machine learning has the advantage of not needing any assumptions about the distribution of the data since the algorithm will learn from the data and find links between variables which might take a long time to detect otherwise.

The main objective of both classification and regression trees is to recursively find a sufficiently good partition of the original data where the relationship between the explanatory variables is made simple and possibly linear models can be applied to accurately predict the response variable. The theory behind the concepts in this subsection can be found in Tan *et al.* (2006).

The input data for a regression tree is a set of records or instances, composed of values from the explanatory variables and values from the response variable to be predicted. Formally, the tree structure is a graph, an hierarchical structure consisting of three types of nodes and directed edges:

- A **root node**, the starting point of the structure, with no incoming edges and zero (when the tree has just one node) or more outgoing edges;
- **Internal nodes**, with exactly one incoming edge and two or more outgoing edges;
- **Leafs** or terminal nodes, with exactly one incoming edge and no outgoing edges, containing the outputs of the structure.

All but the terminal nodes contain explanatory variable test conditions to separate records that have different characteristics. Furthermore, there is a single model attached to each leaf. If it is a decision tree, then the output will be a class of the response variable, usually the majority one. If it is a regression tree, as in our study, then the output will be by default the average value of the observed response variable for that leaf.

Given this framework, there is an exponential number of different trees to partition

the same set of explanatory variables, some of which will yield better results than others. Given the number of possible trees, finding the optimal one is computationally infeasible. Numerous algorithms for finding sub-optimal but reasonably accurate options have been developed, hinging on two key aspects:

- **How the records should be split:** Finding the right test condition to partition the data which goes to each particular node.
- **Stoping the splitting procedure:** Finding the point at which it is no longer beneficial to grow the tree. Several types of stoping conditions can be used: maximum number of nodes in the tree; minimum number of records in a new node, *etc.*

Three types of regression trees have been chosen for this study, for their simplicity and good prediction accuracy: CART, conditional inference trees and random forest. These will be briefly introduced in the following sub-subsections. Further detail can be found for instance in Tan *et al.* (2006); Murthy (1998) and Safavian and Landgrebe (1991).

2.2.5.1 Classification and regression trees (CART)

CART is a classification and regression tree algorithm, first presented in Breiman *et al.* (1984). This method uses binary splits, i.e., all explanatory variable tests divide the data at each node into two subsections. Binary splits are not mandatory for regression trees but are common.

The splitting criteria is the combination of left and right nodes which maximises $SS_P - (SS_L + SS_R)$, where P represents the parent node, L its left son and R its right son and $SS_{node} = \sum(y_i - \overline{y})^2$. This is equivalent to choosing the split to maximise the between-groups sum-of-squares in an analysis of variance.

To stop the splitting process, CART follows a strategy where a tree as big as possible is grown, stoping only when some minimum node size is reached. Afterwards, the tree is pruned back to a smaller size. Pruning is a technique used for reducing the size of trees by removing sections of the tree that have little power to predict new instances thus avoiding overfitting.

Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. This usually happens when a model is excessively complex, such as having too many parameters or, in the case of trees, too many nodes. A model that is overfit will generally have poor predictive performance.

In CART, a cost-complexity based methodology is employed using cross-validation (an approach in which each record of a dataset is used the same number of times for training and exactly once for testing).

2.2.5.2 Conditional inference trees

Conditional inference trees are a kind of tree, proposed by Hothorn *et al.* (2006), designed to solve the problem of bias towards variables with many possible splits when doing recurs-

ive partitioning. It is an algorithm based on statistical properties of the variables which grows trees that avoid both the overfitting and the variable selection problems.

Firstly, the algorithm tests the global null hypothesis of independence between any of the explanatory variables and the response variable. It stops if this hypothesis cannot be rejected. Otherwise, it selects the explanatory variable with strongest association to the response variable (which is measured by a p-value corresponding to a test for the partial null hypothesis of a single explanatory variable and the response variable).

Secondly, a binary split is implemented in the selected explanatory variable. These two steps are then repeated recursively, until a statistical criteria to stop the algorithm is reached, preventing overfitting without resorting to pruning. This algorithm keeps the processes of choosing and splitting of the explanatory variables separate in order to avoid the bias problem.

2.2.5.3 Random forests

Introduced by Breiman (2001), a random forest is a type of ensemble method, meaning it combines several predictors in order to create a single one that surpasses any of its parts. In a random forest, instead of choosing the best split among all input variables, each node is split using the best among a subset of predictors randomly chosen at that node. This strategy performs quite well compared to many other classifiers and is robust against overfitting.

Roughly, the algorithm first draws n bootstrap samples of the original data, where n is the number of trees in the forest. Then, for each of these, it grows an unpruned regression tree, for which it randomly samples m of the predictors and chooses the best split from those. Lastly, it makes predictions by calculating the average of the predictions of its n trees.

However, with this algorithm, the graphical explainability of having just one tree is lost since random forests can be applied with hundreds of trees as the basis.

2.2.6 Hybrid models

For this work, the main advantage of regression trees is partitioning the data, as inspired by Guo *et al.* (2002). Several characteristics of the insured lives in our study were available and using trees to partition these characteristics was very useful.

However, since we'll be modelling the probability of death of an insured person and regression trees give one result for each subset of the data (the average over that subset), we will go beyond using the predictions from the trees. As a new approach, the models from Subsections 2.2.1 (Gompertz's law), 2.2.2 (empirical approach) and 2.2.4 (GLM) will be used in combination with CART and conditional inference trees. Random forest is not used in combination with other methods because, since it generates a large number

of trees, it would be computationally too expensive. Details of this implementations are given in Section 4.6.

2.2.7 Model evaluation

Before explaining the metric used for model evaluation, the meaning of some specific concepts must be clarified. More detailed descriptions of what follows can be found in Tan *et al.* (2006); Hastie *et al.* (2001).

To evaluate the performance of the various approaches, all models were applied after dividing the data into a training and a test set. To divide the data, we chose to use the holdout method, in which the original data set is divided into two disjoint subsets - one for training, one for testing. The first subset is used for defining the model (its parameters, in the case of parametric models) and the second subset is used only for measuring accuracy (by running the data through the model and measuring how different the predicted values are from the true values). This allows the user to get an unbiased estimate of the accuracy of the method in question. Other techniques, such as cross-validation or bootstrapping (where records are sampled with replacement, i.e., a record already chosen is put back into the original pool of records so that it is equally likely to be redrawn) could have been used.

The procedure of dividing the data, defining the model using the training set and evaluating it on the test set is meant to get an independent evaluation of the model by removing the possibility of it being overfit to the training set.

For a performance metric we needed one which, for a previously unused subset of the original data (the test set), compared predicted values with the actual values and quantified the differences between them. We will use the RMSE (root mean square error) as the measure of performance and the tool to compare the various methods. This measure is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}, \tag{2.2.4}$$

where y_i is the real value of the i-th object, $\hat{y}_i$ its estimated value and n is the sample size. It compares the difference between predicted and actual values in a smooth way.

Chapter 3

Exploratory Data Analysis

3.1 The data base

In this section, we will describe the data in study and the cleaning processes that were applied, which were done using the software SAS Enterprise Guide.

The data comes from an insurance company's life risk group portfolio from 2007 until 2014. It is composed of information retrieved at the end of each year (31/12/2007 to 2014) and March 2015. This information was then organised with one line per life in each policy (a policy which covers two lives will have two lines), which originated a total of 1.942.581 records. Each record contained the following relevant fields:

- ID- a unique identifier of each record, containing policy number;
- Person key - a unique identifier of each person insured over all the database;
- Gender - gender of the insured person (male or female);
- Date birth - date of birth of the insured person;
- Date issue - date from which the person was insured;
- Contract status - contract status on 31/03/2015 - in force, term (ended on end date), lapse (ended before the term), claim (ended due to a claim);
- End date - date at which the person stopped being insured, if the contract is no longer in force (empty if contract status is "in force");
- SA 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014 - respectively, sum assured on 31/12/2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014 (in €);
- Death claim - one if there was a death claim, zero if there wasn't a death claim or it was rejected;
- Claim cost - if there was a claim, how much it cost (in €);
- Civil status - single, married, widowed or divorced;
- District - district of the address on record for the insured person (Porto, Lisboa, Setúbal, Açores, Viseu, Santarém, Évora, Coimbra, Faro, Vila Real, Leiria, Beja, Aveiro, Braga, Bragança, Guarda, Viana do Castelo, Madeira, Castelo Branco, Portalegre or "Outside of Portugal"). These are not the official districts of Portugal, the island territories (Ilhas) were grouped due to the lack of records.

In order to clean the database, records which verified the following criteria were excluded:

- No gender, birth, start date, end date (when not in force), civil status or district;
- District was "Outside of Portugal";
- The person was younger than sixteen at start date;
- The start date was posterior to the end date.

After these exclusions, fields date of birth, gender, civil status and district were analysed and corrected in reference to the person key field due to their importance for this study. The person key field should yield a unique identifier of each different person in the database. From the 1.207.639 different person keys in the database, 0,34% had more than one combination of (gender, date of birth, civil status, district). To mitigate this, the following steps were taken, followed by further exclusions:

- If there was one most common combination (higher frequency and unique) for that key where the contracts were in force, that combination was chosen;
- Otherwise, if there was one most common combination (higher frequency and unique) for that key (regardless of the contract status), that combination was chosen;
- Otherwise, nothing was done.
- Insured lives with more than one combination of (gender, date of birth, civil status, district) - the cases that couldn't be corrected - were excluded;
- Contracts without a sum assured when they were in force were excluded;
- Contracts where the date of issue was equal to the end date were excluded.

This accounted for 707.815 less records in the data. In the end, 1.234.664 records were left, representing 781.772 insured lives. A decision was made to analyse the period between 01/01/2011 and 31/12/2014 and so contracts which weren't in force during that period were excluded and the Date of Issue and End Date fields were modified to simplify calculations to come, since exposure to risk outside of the analysis period is irrelevant to the study.

- If date of issue was before 01/01/2011, it was considered to be 01/01/2011;
- If end date was after 31/12/2014, it was considered to be 01/01/2015. This includes contracts still in force at the end of the analysis period.

Upon analysing the district variable, we realised the multitude of possible values (twenty) made it too hard to draw conclusions. As such, the field NUTS_M was constructed as a proxy of the NUTS - *Nomenclatura Comum das Unidades Territoriais Estatísticas* - II for Portugal, whose definition can be found in (Instituto Nacional de Estatística 2013). The new field NUTS_M has the following six possible values: Ilhas (districts Madeira and Açores), Centro (districts Aveiro, Castelo Branco, Coimbra, Guarda, Leiria, Santarém and Viseu), Alentejo (districts Beja, Évora and Portalegre), Norte (districts Braga, Bragança, Porto, Viana do Castelo and Vila Real), Algarve (district Faro) and Área Metropolitana Lisboa (districts Lisboa and Setúbal).

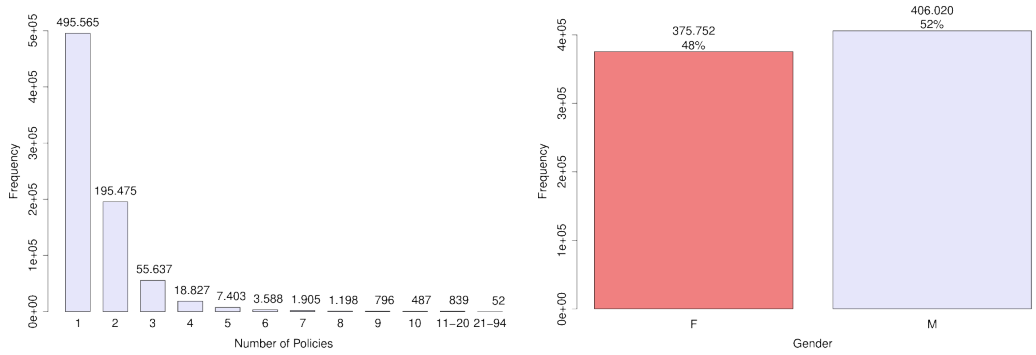
For contracts where death occurred, the sum assured on the year of death and the claim cost were compared. The absolute difference between these two amounts was computed for each claim as $abs(SA\_Claim - ClaimCost)$, where SA_Claim is the sum assured on the

year of death. Upon inspection, there is an average absolute difference of 880.2€ between the two values and only 85.75% of the claims had no difference. This could happen for a number of reasons but since these differences were not the point of this work, we will from now on assume that the claim cost was equal to the sum assured on the year of death, when there was a claim.

3.2 Data analysis

In this section, the fields from the previous section will be analysed, using various types of graphics and descriptive measures. The computational work was performed using the software R.

- Person key - Since this field is used as a unique identifier of the insured lives, we used it to count the number of policies each person has in the portfolio. The results are presented in the bar plot in Figure 3.1 (a). In these, we can see that 63% of people have only one policy between 2011 and 2014. We can also observe there was at least 1 person with 94 policies, which is probably an error in the data.



(a) (b)

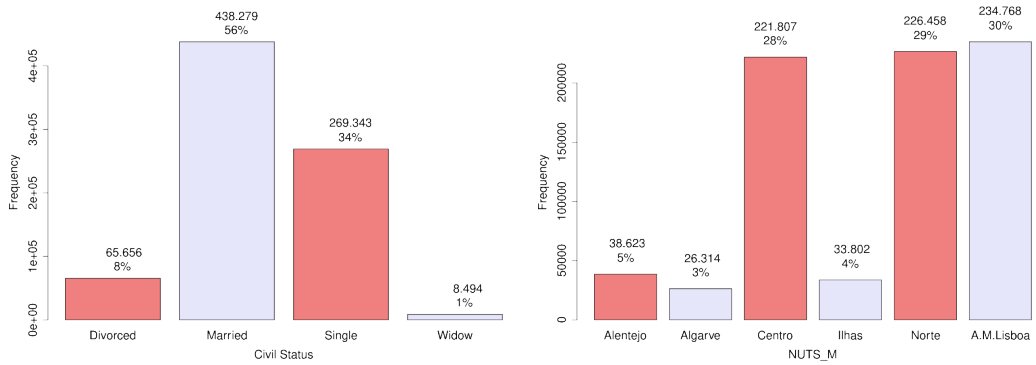
Figure 3.1: (a) Number of policies per insured life; (b) Number of insured lives per gender.

- Personal Information (these variables were analysed per person key, i.e insured life, instead of contract):
  - Gender: gender distribution in the portfolio is illustrated in the bar plot of Figure 3.1 (b). We can observe that the portfolio is close to balanced when it comes to the proportion of males and females.
  - Age: using the date of birth, the real age (the integer part of the exact age) of the insured persons at the end of each civil year of the analysis period was calculated. Table 3.1 shows descriptive statistics for these variables. The portfolio is clearly getting older, as attested by the increase in 1st and 3rd quantiles, median and mean over the years.

Table 3.1: Ages of insured lives throughout the study (in years)

Date	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
31/12/2011	17	35	43	43,56	51	101	11
31/12/2012	17	36	43	44,24	52	102	11
31/12/2013	17	37	44	44,84	52	103	10
31/12/2014	18	38	45	45,45	53	97	10

  - Civil status: In the bar plot of Figure 3.2 (a) we can see that most (around 56%) of the people in this portfolio are married and that the second biggest group is the group of singles.



(a) (b)

Figure 3.2: Number of insured lives per: (a) Civil status; (b) NUTS_M.

  - NUTS_M: The distribution of the insured persons of the portfolio according to their NUTS_M of residence is visible in the bar plot of Figure 3.2 (b). The Lisboa area is the most significant, followed closely by Norte and Centro.

- Death Claim - This variable is zero or one depending on wether the contract in question had a death claim or not. Figure 3.3 shows that there were about 5.000 claims between 2011 and 2014.

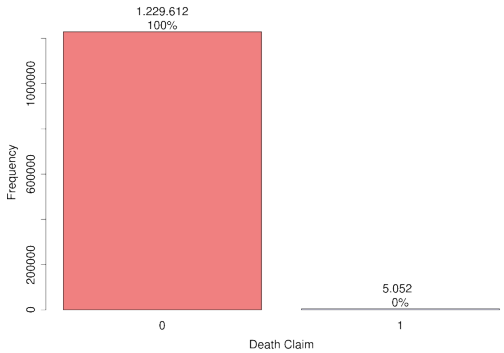


Figure 3.3: Number of death claims

- SA_2011, SA_2012, SA_2013, SA_2014 - These variables represent the sum assured

per policy in force on 31/12/2011, 2012, 2013 and 2014 (in €). Table 3.2 shows descriptive statistics for these variables. We can see that the total sum assured in the portfolio diminished by over seven billion € in the four year period.

Table 3.2: Sum Assured for policies in force throughout the study (in €)

Variable	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation	Total Sum Assured
SA_2011	0,01	11.190	29.820	46.430	65.950	12.530.000	52.358	51×10^9
SA_2012	0,02	11.670	30.110	46.410	65.200	12.530.000	52.106	49×10^9
SA_2013	0,02	12.060	30.920	46.420	64.920	12.530.000	51.405	46×10^9
SA_2014	0,01	12.430	31.350	46.330	64.600	12.530.000	50.438	44×10^9

3.3 Exposed to risk

The concepts from Section 2.1 will now be used as the risk exposure of each person in the portfolio is calculated. The real age of the insured persons and an actual/actual base of calendar will be considered.

First, it was decided that the mortality graduation would be based on the insured lives and not on the policy participants, since one person can have more than one policy. As such, the same person, across all policies where he or she is insured, should only contribute to the risk exposure once per period of time.

Furthermore, the sum assured of each policy will be assumed to have changed at most once a year, with the start of each civil year. As such, our risk exposure calculation should be broken down by civil year, to make it possible to link the right sum assured to each period of time. We will now give an example.

Consider a person who was insured for the whole of 2011 and turned 50 on 01/07/2011. Following equation 2.1.1, this person will have

$$E_{49} = \frac{\#\,days\,exposed\,risk\,(age\,49)}{\#\,days\,year\,(2011)} = \frac{181}{365},\ E_{50} = \frac{\#\,days\,exposed\,risk\,(age\,50)}{\#\,days\,year\,(2011)} = \frac{184}{365}.$$

Consider now that the same person had another policy in force from 01/08/2011 to 31/12/2011. If E_x was calculated without taking into account the person's policies as a whole, we would have arrived to the conclusion that for 2011

$$E_{49} = \frac{181}{365},\ E_{50} = \frac{184+153}{365} = \frac{337}{365}.$$

As such, that person would have counted twice for age 50 during the period of 01/08/2011 to 31/12/2011, which wouldn't be coherent with our previous decision. Since the period in force of the second policy overlaps with that of the first policy, the correct result is

$$E_{49} = \frac{181}{365},\ E_{50} = \frac{184}{365}.$$

As stated in Subsection 2.1.1, we will be studying the mortality rates weighted by sum assured, which requires the weighted risk exposure, given by equation 2.1.3.

Continuing with the previous example, assume the first policy had 2.500€ of sum assured and the second had 50.000€. For the period of 01/01/2011 to 01/08/2011, the sum assured was 2.500€. For the period of 01/08/2011 to 31/12/2011, it was 2.500€ + 50.000€ = 52.500€ since the person was covered by both policies. Hence:

$$E_{49}^{SA} = \frac{181}{365} \times 2.500 = 1.240;\ E_{50}^{SA} = \frac{184}{365} \times 2.500 + \frac{153}{365} \times 50.000 = 22.220.$$

Regarding the number of deaths at a given age x, d_x, considering once again that a single person could have multiple policies and, as such, a death claim in multiple polices, care had to be taken not to count the same person's death more than once when calculating the mortality rates. For the mortality rates weighted by sum assured, the sum of the sums assured over all policies for which a death was reported will be used as d_x^{SA}.

In order to perform the calculations previously explained, an algorithm was created to transform the data into a form where the exact risk exposure and number of deaths for any segment of the portfolio would be easily calculated. The objective was to break a policy's (date of issue, end date) interval into smaller sub-intervals where there were no changes in age or sum assured.

The algorithm is explained in pseudo-code in Appendix B. The result was a data set with 3.379.418 records for the 781.772 different insured lives with the following attributes: person key, gender, civil status, NUTS_M, Ex, age, ExSA, claim, SA_Claim.

3.4 Mortality rates

In this section, we'll look at the crude mortality rates (weighted and not) by age for the portfolio in study, first over all insured lives and then separated by each of the personal explanatory variables available (gender, civil status and NUTS_M).

The individual identifier (person key) was dropped and the variables Ex, ExSA, claim and SA_Claim were summed grouped by the personal characteristics. This produced 2.877 different combinations. In order to keep the results significant, we limited our analysis between ages 20 (before that there were no deaths) and 80 (after that there is barely any exposure to risk), leaving 2.655 different cohorts.

This structure will be the basis for constructing the observed mortality rates. The remainder of this section will be divided into two perspectives, using the data aggregated by age, gender, civil status and NUTS_M:

- Mortality rates: Given by equation 2.1.2, where d_x and E_x are, respectively, the number of claims (field Claim) and the risk exposure (field Ex) for age x.
- Weighted mortality rates: Given by equation 2.1.4, where d_x^{SA} and E_x^{SA} are, respectively, the sum assured of claims (field SA_Claim) and the risk exposure (field ExSA)

for age x.

Throughout this section, both approaches will also be compared via the formula

$$r = mean\left(\frac{\tilde{q}_x^{SA}}{\tilde{q}_x}\right), \tag{3.4.1}$$

where $\tilde{q}_x$ and $\tilde{q}_x^{SA}$ are the observed values of q_x and q_x^{SA} which are bigger than zero (otherwise, the ratio would not be computable). The mean of the ratio is calculated grouped by the explanatory variable in analysis. The results of this value r provide a notion of how much higher (or lower) q_x is than q_x^{SA}, on average.

3.4.1 Overall mortality rates

- Mortality rates: In Figure 3.4 we can see the mortality rates for the portfolio over ages 20 to 80. It resembles a slow growing exponential curve, which picks up at around age 65.
- Weighted mortality rates: In Figure 3.4, it's possible to see the overall mortality rates for all ages weighted by sum assured. The behaviour is similar that of the traditional rates except that the growth with age seems to accelerate later.

Comparing the two approaches (Figure 3.4), it seems like q_x is higher than q_x^{SA} for the most part, especially for older ages. As a matter of fact, calculating r as stated before, q_x^{SA} is on average 86% of q_x.

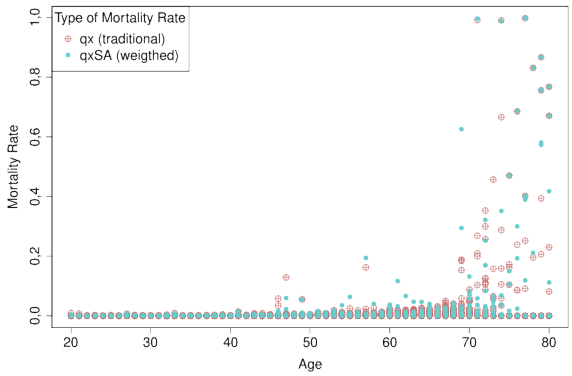
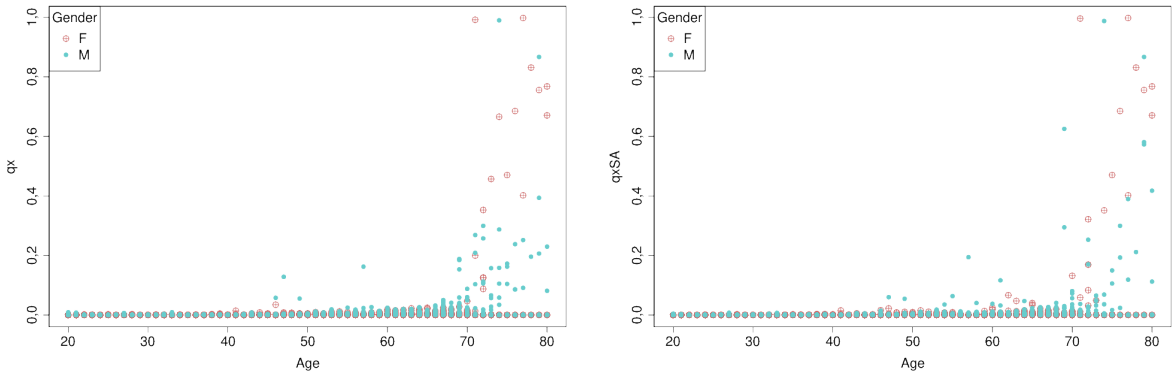


Figure 3.4: Overall mortality rates *versus* mortality rates weighted by sum assured

3.4.2 Mortality rates by gender

- Mortality rates: Figure 3.5 (a) shows the mortality rates for the portfolio for male *versus* female insured lives. From visual inspection of these plots, it seems like the female mortality rates are lower than the males' for younger ages but then become higher after age 70.
- Weighted mortality rates: Analysing the weighted mortality by gender (visible in Figure 3.5 (b)), we can see the same behaviour.

Comparing the two approaches (plots have the same scale), the weighted rates seem lower. Calculating the mean of the ratio between the values (r), q_x^{SA} is on average 93% of q_x for females and 81% for males.



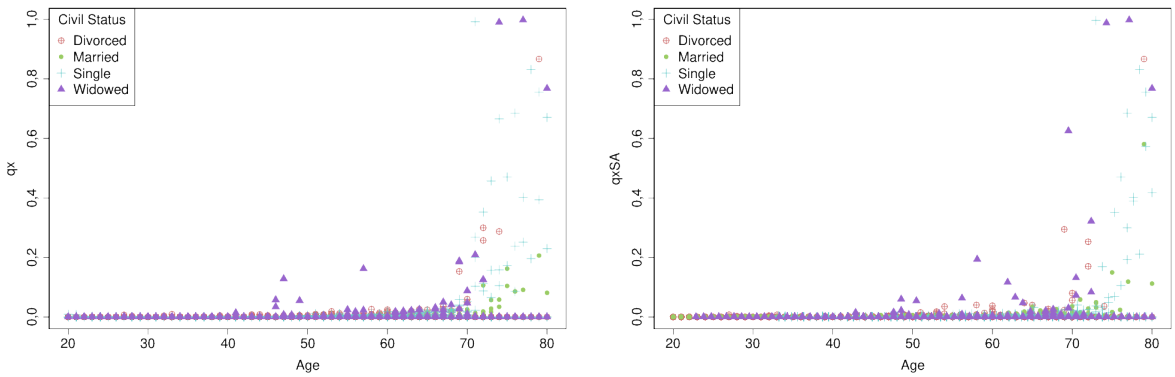
(a) (b)

Figure 3.5: Mortality rates by gender: (a) traditional; (b) weighted.

3.4.3 Mortality rates by civil status

- Mortality rates: Figure 3.6 (a) has the mortality rates divided by civil status. For this segmentation, it's harder to draw conclusions given that there are more possible values for the explanatory variable.
- Weighted mortality rates: Values are plotted in Figure 3.6 (b), where once again no clear conclusions are possible.

Comparing the two approaches, it's hard to say visually which one yields higher mortality rates. Calculating the ratio r, q_x^{SA} is on average: 87% of q_x for divorced people; 91% of q_x for married people; 77% of q_x for singles; 108% of q_x for widowed people. In the case of widowed lives, the traditional mortality rates are for the first time higher than its weighted counterpart.



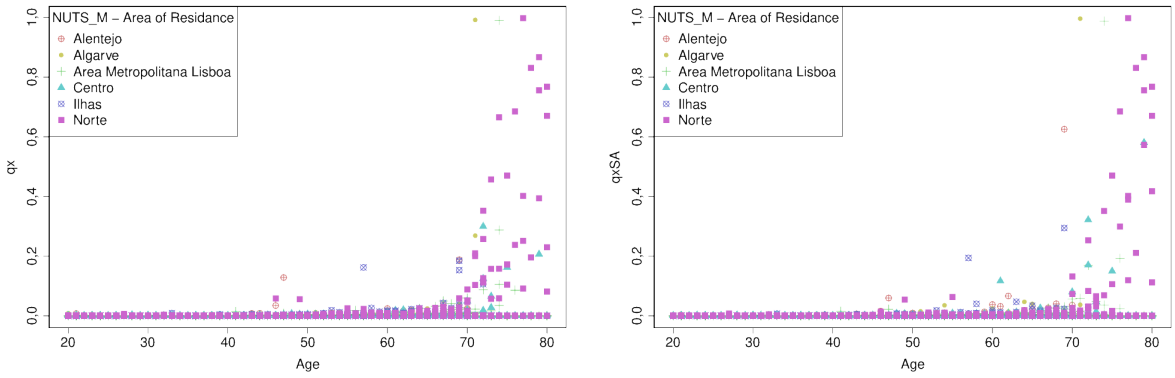
(a) (b)

Figure 3.6: Mortality rates by civil status: (a) traditional; (b) weighted.

3.4.4 Mortality rates by NUTS_M

- Mortality rates: Finally, Figure 3.7 (a) shows the different mortality rates according to NUTS_M of residence.
- Weighted mortality rates: Figure 3.7 (b) has the weighted mortality rates.

Trying to reach conclusions about the two approaches visually is pointless. Comparing the two through the mean ratio r, q_x^{SA} is on average: 100% of q_x for Alentejo; 88% of q_x for Algarve; 86% of q_x for Lisboa; 96% of q_x for Centro; 80% of q_x for Ilhas; 75% of q_x for Norte.



(a) (b)

Figure 3.7: Mortality rates by NUTS_M: (a) traditional; (b) weighted.

Comparing the values of the mortality rates *versus* the weighted mortality rates throughout, for most segments, q_x^{SA} is lower than q_x, which is in line with the idea that people with higher sums assured are wealthier and hence healthier, leading to longer lives.

3.5 Training and test sets

As discussed in Section 2.2, a training and a test set were used. The training set was composed of 80% of the insured lives in the data set (624.880) and the remaining 20% (156.220) as the test set. Note that this division was made over the number of unique insured lives and not the number of policyholders, so as to be coherent with the exposure and claim calculations.

In order to have a similar composition over the available explanatory variables of the training and test sets, compared to the original data set, stratified sampling was used, instead of directly doing a random sample of 80% of the insured lives. This is meant to avoid the risk of over-representing one segment of the population in the training set and under-representing it in the test set, or *vice-versa*.

The method of stratification used is described in pseudo-code in Appendix C. This procedure assured that the training and test sets have the same percentage as the original data of each stratum (gender, civil status, NUTS_M) regarding number of unique people.

Chapter 4

Results

In this chapter, the models of Section 2.2 will be applied to the training set. In the first four sections, the details and results of the application of each model will be presented, always divided by the two approaches of this thesis: (traditional) mortality rates and mortality rates weighted by sum assured. In the last Section (4.7), the models will be compared between them, using the test set to calculate the error.

Furthermore, recall r, as defined in equation 3.4.1. The results of this value will provide an idea of how much higher (or lower) the estimated q_x is than the estimated q_x^{SA}, on average, for each model. Given that for some of the models the traditional and weighted mortality rates are grouped in different ways, the ratio r is only comparable for the overall (ungrouped) rates. These calculated values will be comparable to the ratio for the crude rates in Subsection 3.4.1, which was 86%.

4.1 Gompertz's law

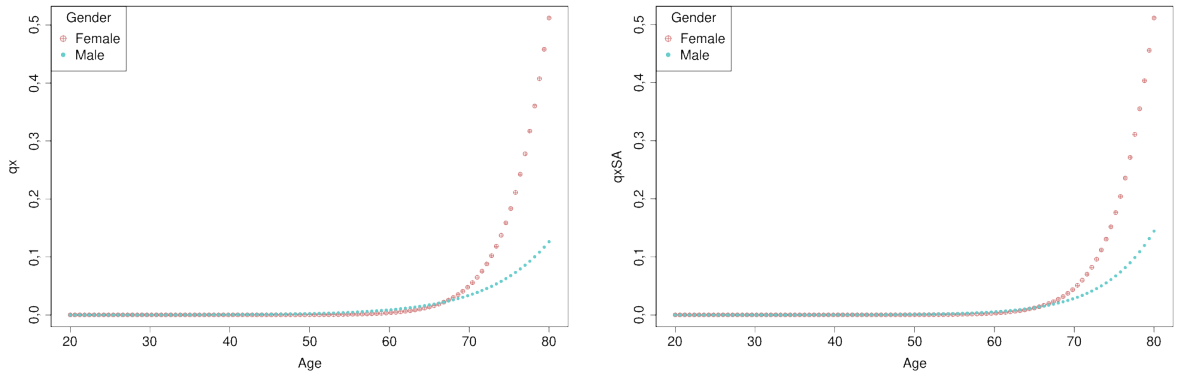
As stated in Subsection 2.2.1, the observed mortality rates were fit to the curve in equation 2.2.1. The two parameters in the equation were determined using the nls function from R, which calculates the nonlinear (weighted) least-squares estimates of the parameters of a nonlinear model. The fitting was done separately for male and female insured lives.

- Mortality rates: The fitting resulted in the curves bellow, visible in Figure 4.1 (a). Looking at the plot of the curves, one can see that the female curve grows much faster than the male curve for ages over 70, like for the crude mortality rates (Figure 3.5 (a)).

$$\hat{q}_x = \begin{cases} 1 - exp\left(-5,08 \times 10^{-10} \times exp(0,26\,x)\right), & Gender = female \\ 1 - exp\left(-2,52 \times 10^{-6} \times exp(0,14\,x)\right), & Gender = male \end{cases}$$

- Weighted mortality rates: The final curves are given by the formula bellow, plotted in Figure 4.1 (b). Again, replicating the behaviour of the crude mortality rates (Figure 3.5 (b)), the female curve exceeds the male curve for ages over 70.

$$\hat{q}_x^{SA} = \begin{cases} 1 - exp\left(-2,49 \times 10^{-10} \times exp(0,27\,x)\right), & Gender = female \\ 1 - exp\left(-15,32 \times 10^{-6} \times exp(0,17\,x)\right), & Gender = male \end{cases}$$



(a) (b)

Figure 4.1: Fitted curves for mortality rates- Gompertz's law: (a) traditional; (b) weighted.

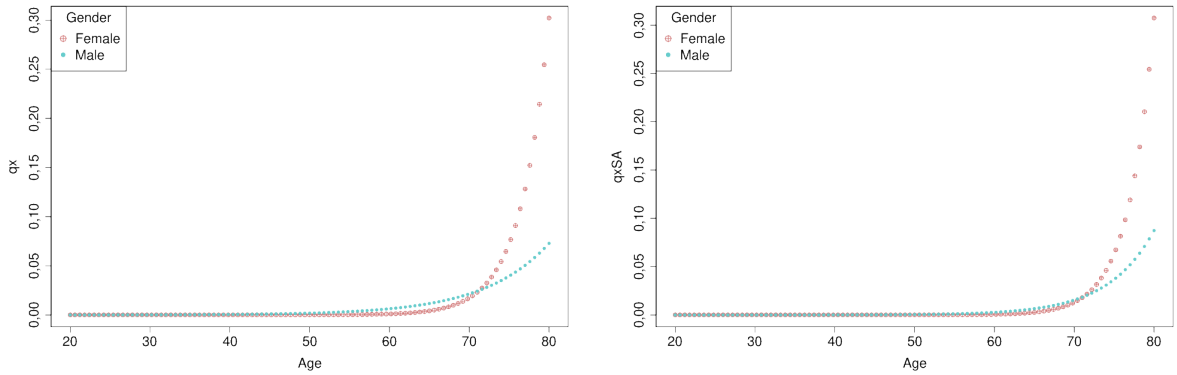
For this model, $\hat{q}_x^{SA}$ is on average 62% of $\hat{q}_x$.

4.2 Empirical approach

By this approach, the observed mortality rates were fit to the curve in equation 2.2.2. Again, the fitting was done separately for male and female insured lives and the parameters were determined using the nls function from R.

- Mortality rates: The fitting resulted in the curves in Figure 4.2 (a). From visual inspection, the male and female curves are indistinguishable up to around age 50. At this point, the male curve starts to exceed the female and the gap grows bigger up to age 70, where the curves intersect again and from that point on the female rates start to surpass the male ones.

$$\hat{q}_x = \begin{cases} 3,57 \times 10^{-10} \times exp(0,29\,x), & Gender = female \\ 4,06 \times 10^{-10} \times exp(0,12\,x), & Gender = male \end{cases}$$



(a) (b)

Figure 4.2: Fitted curves for mortality rates- empirical approach: (a) traditional; (b) weighted.

- Weighted mortality rates: The fitted curves are given by the formula bellow, visible

in Figure 4.2 (b). Looking at the plot, the maximum value for this approach is reached at around 0,3 whereas for the previous model (Gompertz's law), it was 0,5.

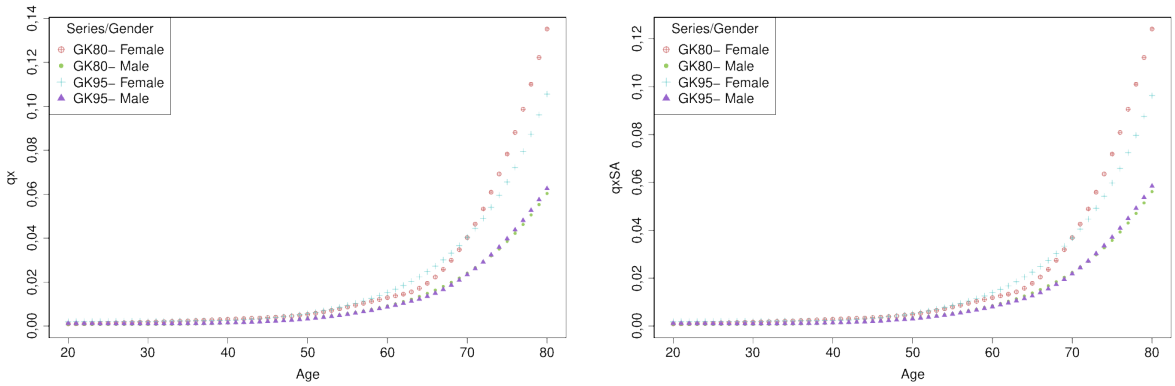
$$\hat{q}_x^{SA} = \begin{cases} 3,20 \times 10^{-12} \times exp(0,32\,x), & Gender = female \\ 8,35 \times 10^{-10} \times exp(0,17\,x), & Gender = male \end{cases}$$

For this approach, the estimated weighted mortality rates are on average 39% of the estimated traditional mortality rates.

4.3 Standard tables

The graduation using standard tables, as stated in Subsection 2.2.3, made use of the Swiss tables GKF80, GKM80, GKF95 and GKM95. These tables were adjusted to the data by gender, i.e., female data was adjusted to GKF80 and GKF95 and male data to GKM80 and GKM95. The formula in equation 2.2.3 was used and the function lm from R, which fits linear models, was utilised to find the parameter.

- Mortality rates: The application of this model resulted in (Figure 4.3 (a)):
  - Using the 80 series: $\hat{q}_x = \begin{cases} 2,82 \times GKF80_x, & Gender = female \\ 0,56 \times GKM80_x, & Gender = male \end{cases}$;
  - Using the 95 series: $\hat{q}_x = \begin{cases} 1,87 \times GKF95_x, & Gender = female \\ 0,74 \times GKM95_x, & Gender = male \end{cases}$.



(a) (b)

Figure 4.3: Fitted curves for mortality rates- standard tables: (a) traditional; (b) weighted.

The high value of the parameter for female data is due to the high values of the observed mortality rates for females over 70 in our data (see Figure 3.6 (a)). These are mostly outliers from a specific policy that existed for a year and was not renewed because of its unprofitable results. As a matter of fact, running the fitting process over the training lives under 70 years old, the results become:

- Using the 80 series: $\hat{q}_x = \begin{cases} 0,34 \times GKF80_x, & Gender = female \\ 0,34 \times GKM80_x, & Gender = male \end{cases}$;

  - Using the 95 series: $\hat{q}_x = \begin{cases} 0,20 \times GKF95_x, & Gender = female \\ 0,48 \times GKM95_x, & Gender = male \end{cases}$.

- Weighted mortality rates: Graduation using the GK80 and GK95 standard tables resulted in the formulas bellow and Figure 4.3 (b).
  - Using the 80 series: $\hat{q}_x^{SA} = \begin{cases} 2,59 \times GKF80_x, & Gender = female \\ 0,55 \times GKM80_x, & Gender = male \end{cases}$;
  - Using the 95 series: $\hat{q}_x^{SA} = \begin{cases} 1,70 \times GKF95_x, & Gender = female \\ 0,69 \times GKM95_x, & Gender = male \end{cases}$.

  As with the traditional rates, the unusually high value of the parameter fit to the female training data is due to the high weighted mortality rates at advanced ages (as can be observed in Figure 3.6 (b)). Fitting the data to the standard tables for lives under 70, the results become:
  - Using the 80 series: $\hat{q}_x^{SA} = \begin{cases} 0,31 \times GKF80_x, & Gender = female \\ 0,36 \times GKM80_x, & Gender = male \end{cases}$;
  - Using the 95 series: $\hat{q}_x^{SA} = \begin{cases} 0,18 \times GKF95_x, & Gender = female \\ 0,50 \times GKM95_x, & Gender = male \end{cases}$.

For the 80 series, $\hat{q}_x^{SA}$ is on average 93% of $\hat{q}_x$. This is corroborated by the fact that the weighted rates are adjusted by a smaller percentage of the standard tables than the traditional rates. For the 95 series, r is 92%.

4.4 GLM

Considering the definition of GLM given in Subsection 2.2.4 the distribution of Y was assumed to be Bernoulli with mean μ. Furthermore, the complementary log-log link function $\eta = cloglog(\mu) = log(-log(1-\mu))$ was used. Hence, $\mu = 1 - e^{-e^{\eta}}$. Several other link functions were tried before deciding to use the complementary log-log function, which presented the best results for the data in question.

As a first approach, for both the traditional and weighted mortality rates, the four available explanatory variables (age, gender, civil status and NUTS_M) were included, with no interaction between them. From this starting point, considering the p-value of the estimated parameters, adjustments were made to the explanatory variables in order to have statistically significant results. Maximum likelihood estimation was then used to estimate the parameters, using an iteratively reweighted least squares algorithm, through the glm function from the software R (stats package).

- Mortality rates: The final model is given by the formula

$$\hat{q}_x = 1 - exp(-exp(-2,00 + 0,02x + CS + N + G)),$$

  where:

- $CS = \begin{cases} 0, & Civil\ Status = divorced \\ 1,06, & Civil\ Status = married \\ 0,96 & Civil\ Status = single \\ -1,27 & Civil\ Status = widowed \end{cases}$;
- $N = \begin{cases} 0, & NUTS\_M = Lisbon \\ -1,31, & NUTS\_M = Alentejo \\ -0,60, & NUTS\_M \in \{Algarve,\ Ilhas,\ Centro,\ Norte\} \end{cases}$;
- $G = \begin{cases} 0, & Gender = female \\ 0,36, & Gender = male \end{cases}$.

No visual representation of this model is available since there are 24 possible combinations of the explanatory variables gender, civil status and NUTS_M.

- Weighted mortality rates: Coincidently, the exact same model was achieved for the weighted mortality rates as for the traditional rates. As such, $\hat{q}_x^{SA} = \hat{q}_x$ and the equations are omitted. No particular reason was found as to why the models coincided.

Given that the models coincided, r is obviously 100%.

4.5 Regression trees

In this section, models explained in Subsection 2.2.5 will be applied. However, unlike the previous applications in this chapter, for the tree generating algorithms, the explanatory variable age was not included because initial tests showed that it was such an important variable when modelling mortality that the first split was always done with it and at around age 70. This lead to a large majority of the data set being grouped into the same branch, which wasn't a useful conclusion. As such, following Guo *et al.* (2002), the insured lives' ages were not included to generate the trees. Age is is included afterwards to model the mortality of the persons that fit in each node of the generated trees, in Section 4.6.

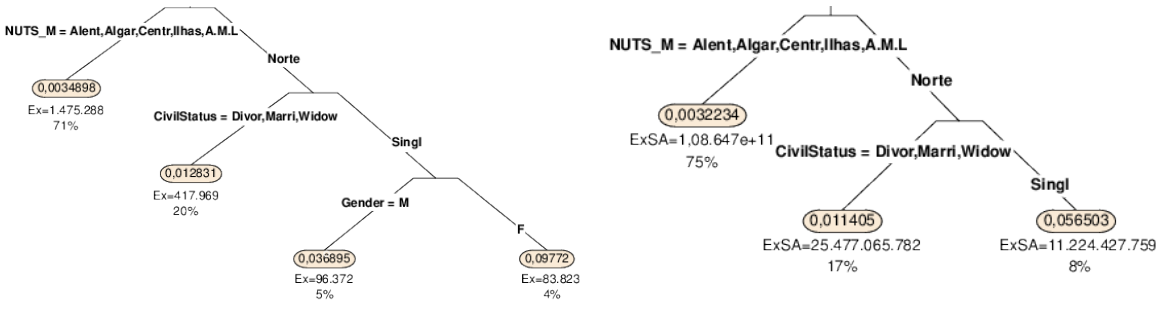
4.5.1 CART

CART (see 2.2.5.1) is the basis for the rpart function from the software R (package rpart), which we used to obtain the results presented bellow. Details of this function can be found in Therneau *et al.* (2015). The minimum node size used was the default, 20.

- Mortality rates: Figure 4.4 (a) shows that the explanatory variable considered most important was the place of residence (NUTS_M), hence being the first split. The remaining explanatory variables are considered important enough to differentiate only those living in Norte. Civil status is divided between singles and the remaining values and male and female lives insured are differentiated for singles from Norte.

Looking at the results from the tree, one might think that despite the age not being provided as an explanatory variable to the algorithm, the choices it made were conditioned by the age of the population that fell within each sub-division, acting as a proxy. However, if we calculate the average age within each leaf (weighted by the risk exposure and for the training data) the results are (from left to right hand-side leafs): 44, 46, 38 and 38 years old. As such, the leaf with the highest average age is the one representing the non-single people from Norte, which has a significantly lower mortality rate estimate than that for the single women from Norte, who are on average 8 years younger.

In Figures 4.4 and 4.5 the shaded boxes have the estimates for the mortality rates of each leaf, bellow each is the total risk exposure of the insured lives in the node (in the training set) and its weight on the overall risk exposure of the training set.



(a) (b)

Figure 4.4: Fitted model for mortality rates - CART: (a) traditional; (b) weighted.

- Weighted mortality rates: Figure 4.4 (b) shows that the results for the weighted and traditional mortality rates using CART are the same, in terms of explanatory variable splits, except that for the weighted mortality rates the gender was not considered important enough.

  As with the traditional mortality rates, we calculated the average age within each leaf (weighted by the weighted risk exposure and for the training data) and the results are (from left to right hand-side leafs): 42, 44 and 37 years old. Again, we can then see that there is no evidence that the algorithm is using the available explanatory variables as a proxy for the age.

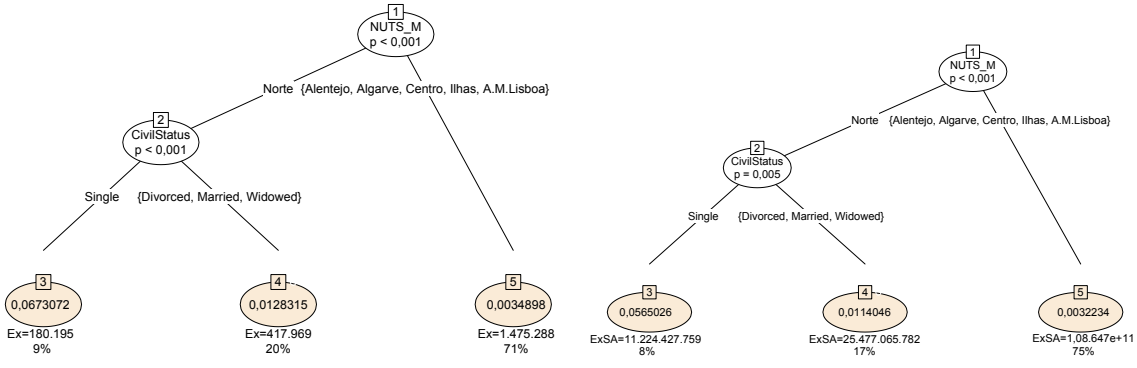
As would be expected, the predicted values for the weighted rates are lower than for the traditional ones. As a matter of fact, for CART, $\hat{q}_x^{SA}$ is on average 93% of $\hat{q}_x$.

4.5.2 Conditional inference trees

To implement this model, we used the R function ctree, from the package party, whose details can be found in Hothorn *et al.* (2014) (see Sub-section 2.2.5.2). The results from this approach are:

- Mortality rates: Figure 4.5 (a) shows the tree grown by the conditional inference trees algorithm. Once again, mortality for Norte is considered to be significantly different from the remaining values of NUTS_M. As with CART, for insured lives from Norte, the algorithm goes on to divide between singles and the remaining civil status. It does not however consider gender important enough to differentiate the estimates. Notice that for nodes where splits are made (using the explanatory variables civil status and NUTS_M) independence tests from the response variable (q_x) had p-values under 0,001.

  Looking at the average weighted age per leaf (for the training data), this will be the same as in the tree generated by CART except that the last split doesn't occur but since the average age was 38 in both leafs, the average for that part of the population will also be 38. As such, the weighted average is (from left to right): 44, 46 and 38 years old. Again, no evidence was found that the algorithm was using the available explanatory variables as a proxy for age.



(a) (b)

Figure 4.5: Fitted model for mortality rates- conditional inference tree: (a) traditional; (b) weighted.

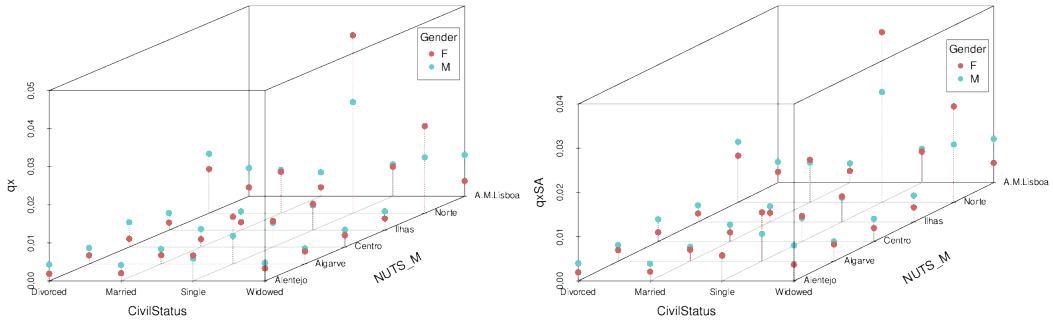
- Weighted mortality rates: Figure 4.5 (b) shows the tree grown for the weighted mortality rates. The splits made are the same as for the traditional mortality rates. In this tree, the NUTS_M explanatory variable's split is done due to the independence test from the response variable (q_x^{SA}) having p-value under 0,001 and for the civil status explanatory variable the p-value is 0,005. Though enough to reject the hypothesis of independence and create a split, it indicates that, according to this algorithm, there is a stronger link between the traditional mortality rates and civil status than between the weighted mortality rates and the same explanatory variable. In this case, the average weighted age per leaf will be exactly the same as with CART (42, 44 and 37 years old, from left to right) and we arrive at the same conclusion that the other variables are not being used as a proxy for age.

Once again, the mortality rate estimates are lower for the weighted version than for the traditional one. Specifically, $\hat{q}_x^{SA}$ is on average 92% of $\hat{q}_x$.

4.5.3 Random forests

To implement the random forests algorithm (explained in Subsection 2.2.5.3), the randomForest function from R (package randomForest) was used, details of which can be found in Liaw and Wiener (2002). For this model, the default number of trees for the randomForest function was used: 500. Tests were made regarding generating up to 100.000 trees but the results were not significantly different.

- Mortality rates: Figure 4.6 (a) shows the results of the Random Forest algorithm for the training set. The plot has only 48 points because this is how many different combinations of (civil status, NUTS_M, gender) there are in the training data. As before, age is not taken into account, so for each combination of these characteristics, there is an unique estimated value. Looking at the plot, one sees that the male rates are most of the times higher than the female rates, which is to be expected. However, the inverse occurs and is exceptionally visible for the single people from Norte. This is in line with the conclusions from the other tree generating algorithms.



(a) (b)

Figure 4.6: Estimated mortality rates using the random forest algorithm: (a) traditional; (b) weighted.

- Weighted mortality rates: Figure 4.6 (b) shows the estimated weighted mortality rates for the training set, which exhibits the same overall behaviour as the traditional rates, although at a smaller scale (here, the maximum vale for the rates is at around 0,04 whereas for the traditional rates it's at around 0,05).

For this approach, the estimated weighted mortality rates are on average 89% of of the estimated traditional mortality rates.

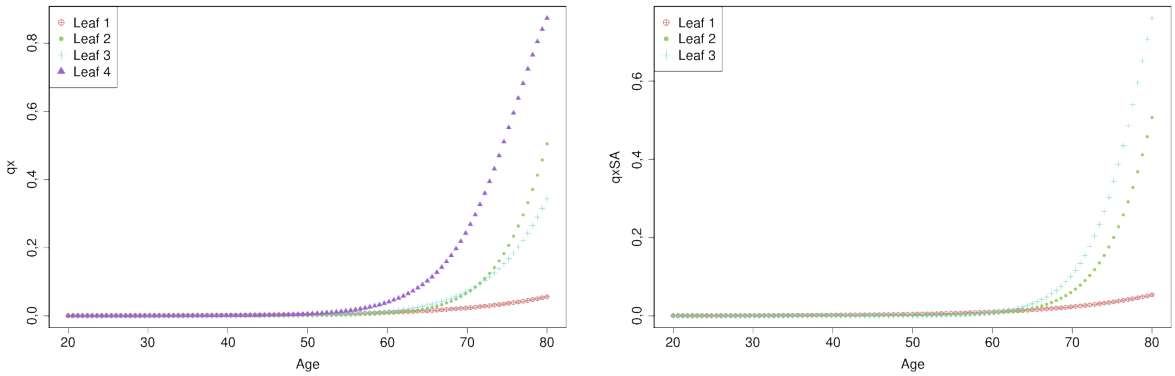
4.6 Hybrid models

Following Guo *et al.* (2002), the tree generating algorithms CART and conditional inference trees were used combined with other methods, namely Gompertz's law, empirical approach and GLM. In this subsection, we will use N for the explanatory variable NUTS_M, CS for civil status and G for gender.

4.6.1 CART and Gompertz's law

- Mortality rates: Recalling Figure 4.4 (a), CART generated a tree with four leafs. Fitting Gompertz's law to the force of mortality of the lives in each leaf results in the curves bellow, Figure 4.7 (a). Looking at the plot, one seems that the curve with the the highest rates is the one corresponding to leaf 4, which is in line with the results from CART.

$$\hat{q}_x = \begin{cases} 1 - exp\left(-4,49 \times 10^{-5} \times exp(0,09\,x)\right), & leaf = 1\,(N \neq Norte) \\ 1 - exp\left(-6,55 \times 10^{-9} \times exp(0,23\,x)\right), & leaf = 2\,(N = Norte \wedge CS \neq single) \\ 1 - exp\left(-4,26 \times 10^{-7} \times exp(0,17\,x)\right), & leaf = 3\,(N = Norte \wedge CS = single \wedge G = male) \\ 1 - exp\left(-2,98 \times 10^{-7} \times exp(0,20\,x)\right), & leaf = 4\,(N = Norte \wedge CS = single \wedge G = female) \end{cases}$$



(a) (b)

Figure 4.7: Fitted curves for mortality rates in each leaf- CART and Gompertz's law: (a) traditional; (b) weighted.

- Weighted mortality rates: According to Figure 4.4 (b), CART generated a tree with three leafs. Fitting Gompertz's law to the force of mortality of the lives in each leaf results in the curves bellow, Figure 4.7 (b). Visual inspection of the plot shows that the curves with lowest to highest mortality rates (1 to 3) are in the same sequence as the lowest to highest estimates in the tree.

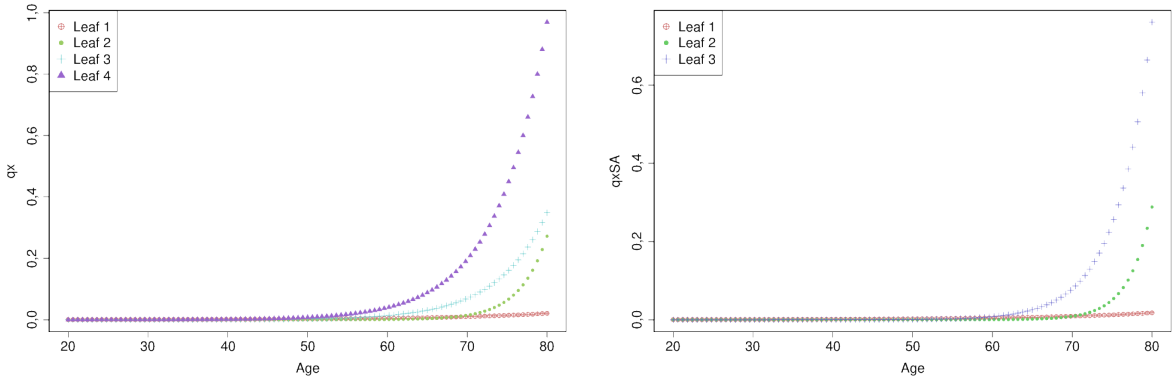
$$\hat{q}_x^{SA} = \begin{cases} 1 - exp\left(-5,68 \times 10^{-5} \times exp(0,09\,x)\right), & leaf = 1\,(N \neq Norte) \\ 1 - exp\left(-3,25 \times 10^{-9} \times exp(0,24\,x)\right), & leaf = 2\,(N = Norte \wedge CS \neq single) \\ 1 - exp\left(-1,96 \times 10^{-7} \times exp(0,26\,x)\right), & leaf = 3\,(N = Norte \wedge CS = single) \end{cases}$$

For this approach, $\hat{q}_x^{SA}$ is on average 100% of $\hat{q}_x$. The ratio between the estimates is 100% only on average, it fluctuates point-wise.

4.6.2 CART and empirical approach

- Mortality rates: Fitting an exponential curve to the mortality rates of the lives in each of the four leafs identified by CART results in the curves bellow and the plot of Figure 4.8 (a).

$$\hat{q}_x = \begin{cases} 6,07 \times 10^{-5} \times exp(0,07\,x), & leaf = 1\,(N \neq Norte) \\ 2,15 \times 10^{-11} \times exp(0,29\,x), & leaf = 2\,(N = Norte\ \wedge CS \neq single) \\ 8,96 \times 10^{-7} \times exp(0,16\,x), & leaf = 3\,(N = Norte\ \wedge\ CS = single\ \wedge\ G = male) \\ 2,71 \times 10^{-6} \times exp(0,16\,x), & leaf = 4\,(N = Norte\ \wedge\ CS = single\ \wedge\ G = female) \end{cases}$$



(a) (b)

Figure 4.8: Fitted curves for mortality rates in each leaf- CART and empirical approach: (a) traditional; (b) weighted.

- Weighted mortality rates: The final model is given by the formula bellow, see Figure 4.8 (b).

$$\hat{q}_x^{SA} = \begin{cases} 7,78 \times 10^{-5} \times exp(0,07\,x), & leaf = 1\,(N \neq Norte) \\ 2,40 \times 10^{-13} \times exp(0,35\,x), & leaf = 2\,(N = Norte\ \wedge CS \neq single) \\ 1,02 \times 10^{-8} \times exp(0,23\,x), & leaf = 3\,(N = Norte\ \wedge\ CS = single) \end{cases}$$

Like in the previous subsection, the sequence of lowest to highest mortality rates estimates (both traditional and weighted) in the tree is the same as in the hybrid models. For this model, $\hat{q}_x^{SA}$ is on average 88% of $\hat{q}_x$.

4.6.3 CART and GLM

For this hybrid model, instead of fitting a GLM model to each of the leafs of the tree, a similar approach to Guo *et al.* (2002) was adopted, using only the first split of the tree to divide the insured lives and then the remaining splits to determine which variables would go into the GLM and how. After generating a first model in this manner, the p-value of the estimated parameters was observed and adjustments were made in order to have statistically significant results.

- Mortality rates: Even though CART has a split where the explanatory variable gender is used (Figure 4.4), upon creating the GLM this explanatory variable had a p-value pointing to no statistical significance. This lead to the explanatory variable not being used in the final model, presented bellow and in Figure 4.9.

$$\hat{q}_x = \begin{cases} 1 - exp(-exp(-2,24 + 0,03x + CS)), & NUTS\_M = Norte \\ 1 - exp(-exp(-2,00 + 0,02x)), & NUTS\_M \neq Norte \end{cases},$$

$$\text{where } CS = \begin{cases} 0, & Civil\ Status \in \{divorced,\ widowed,\ married\} \\ 1,41, & Civil\ Status = single \end{cases}.$$

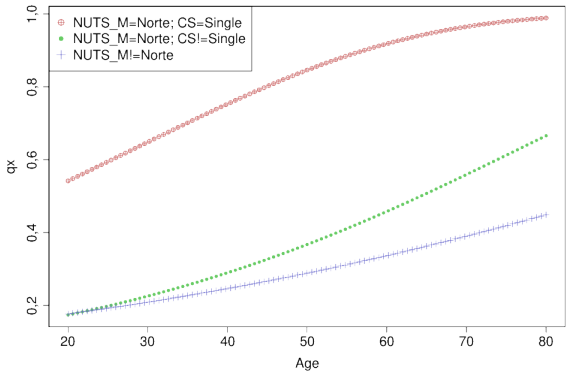


Figure 4.9: Fitted curves for mortality rates- CART and GLM

- Weighted mortality rates: As in Subsection 4.4, the exact same model is achieved for both the weighted mortality rates and the traditional rates. Hence, $\hat{q}_x^{SA} = \hat{q}_x$ and the equations and plot of these are omitted.

For this type of hybrid model, the curves per leaf present a very different behaviour from those of the previous two subsections, especially for the single people from Norte (the red curve), which is convex instead of concave. Still, this segment of the population has the highest mortality rates estimates, in line with the results from the tree. Contrary to the crude mortality rates, for these models there isn't a faster growth after age 70. Given that the models coincided, r is obviously 100%.

4.6.4 Conditional inference trees and Gompertz's law

- Mortality rates: Recalling Figure 4.5, the conditional inference trees algorithm generated a tree with three leafs. Fitting Gompertz's law to to the force of mortality of the lives in each leaf resulted in the curves bellow.

$$\hat{q}_x = \begin{cases} 1 - exp\left(-2,59 \times 10^{-7} \times exp(0,19\,x)\right), & leaf = 1\,(N = Norte\ \wedge CS = single) \\ 1 - exp\left(-6,55 \times 10^{-9} \times exp(0,23\,x)\right), & leaf = 2\,(N = Norte\ \wedge CS \neq single) \\ 1 - exp\left(-4,49 \times 10^{-5} \times exp(0,09\,x)\right), & leaf = 3\,(N \neq Norte) \end{cases}$$

  Naturally, the fitted curves for leafs two and three coincide with the fitted curves for leafs two and one, respectively, for the CART tree since the insured lives inside each of these leafs is the same. Figure 4.10 shows the plot of the three curves, where a rapid growth of the rates is observable for ages over 60 in the case of leafs 1 and 2.

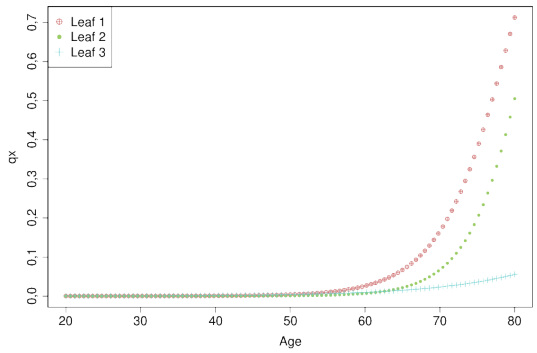


Figure 4.10: Fitted curves for mortality rates in each leaf- conditional inference trees and Gompertz's law

- Weighted mortality rates: Since for the weighted mortality rates the trees generated by the algorithms CART and conditional inference trees coincide, so did this model. Hence, the results are exactly the same as in Subsection 4.6.1.

For this approach, $\hat{q}_x^{SA}$ is on average 100% of $\hat{q}_x$. The ratio between the estimates is 100% only on average, it fluctuates point wise.

4.6.5 Conditional inference trees and empirical approach

- Mortality rates: Fitting an exponential curve to the mortality rates of the lives in each of the four leafs identified by the conditional inference tree algorithm, results in the curves bellow and Figure 4.11

$$\hat{q}_x = \begin{cases} 1,81 \times 10^{-6} \times exp(0,16\,x), & leaf = 1\,(N = Norte\ \wedge CS = single) \\ 2,15 \times 10^{-11} \times exp(0,29\,x), & leaf = 2\,(N = Norte\ \wedge CS \neq single) \\ 6,07 \times 10^{-5} \times exp(0,07\,x), & leaf = 3\,(N \neq Norte) \end{cases}$$

  Again, the fitted curves for leafs two and three of this tree coincide with the curves for leafs two and one, respectively, for the CART tree. For this model, the maximum mortality rate estimate is lower than for the model in the previous section (CIT and Gompertz's law).

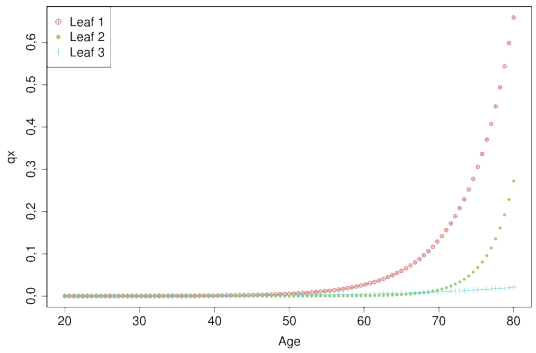


Figure 4.11: Fitted curves for mortality rates in each leaf- conditional inference trees and empirical approach

- Weighted mortality rates: Again, since for the weighted mortality rates the CART and conditional inference trees coincide, so did this model. Hence, the results are exactly the same as in Subsection 4.6.2.

For this model, $\hat{q}_x^{SA}$ is on average 87% of $\hat{q}_x$.

4.6.6 Conditional inference trees and GLM

As in Subsection 4.6.3, only the first split of the tree is used to divide the insured lives and the remaining splits were used to determine which variables go into the GLM.

- Mortality rates: As explained in the Subsection 4.6.3, even tough CART split using the gender explanatory variable, for the GLM, it was not considered statistically significant and removed. This led to the final model being given by the exact same formulas and figure presented in Subsection 4.6.3, not repeated here.
- Weighted mortality rates: Once more, since for the weighted mortality rates the CART and conditional inference trees coincide, so does this model. Hence, the results are exactly the same as in Subsection 4.6.3.

Given that the models coincided, r is 100%.

4.7 Model evaluation

In this section, the RMSE (equation 2.2.4) of each model is calculated and the best results are identified. The RMSE was calculated on the test set, to properly evaluate the accuracy of the models on a data set they had never seen before.

4.7.1 Traditional mortality rates

Table 4.1 summarises the error for each model and the parts of the text related to each of them. Looking at the results, the models using GLM clearly had the worst results (39,6%), all the others having RMSE of at most 5%. The best five results were reached using trees in two of the models and what would be considered more traditional approaches in the

remaining three. The use of the trees combined with Gompertz's law yielded worse results than Gompertz's law alone, whereas either tree with the empirical approach improved on the results of the empirical approach applied alone. The best graduation for the mortality rates was the model for CART combined with the empirical approach.

Table 4.1: Test RMSE per Model (traditional mortality rates)

Model	**Theory**	**Results**	**Test RMSE**	**Rank**
Gompertz's law	2.2.1	4.1	0,0482	8
empirical approach	2.2.2	4.2	0,0465	2
standard table (GKM/F80)	2.2.3	4.3	0,0471	4
standard table (GKM/F95)	2.2.3	4.3	0,0471	5
GLM	2.2.4	4.4	0,3962	13
CART	2.2.5.1	4.5.1	0,0485	9
CIT	2.2.5.2	4.5.2	0,0482	7
RF	2.2.5.3	4.5.3	0,0481	6
CART + Gomp.	2.2.6	4.6.1	0,0514	11
CART + emp.	2.2.6	4.6.2	0,0448	1
CART/CIT + GLM	2.2.6	4.6.3, 4.6.6	0,3455	12
CIT + Gomp.	2.2.6	4.6.4	0,0511	10
CIT + emp.	2.2.6	4.6.5	0,0468	3

4.7.2 Weighted mortality rates

Table 4.2 is similar to Table 4.1, now for the weighted mortality rates. Looking at the results, once again, the models using GLM had the worst results (39,6%), all the others having RMSE bellow 5%. The best five results were reached using trees in one of the models and more traditional approaches in the remaining four. As with the traditional mortality rates, the use of the CART/CIT tree (they coincided) combined with Gompertz's law led to worse results that Gompertz's law alone. Again, the tree with the empirical approach improved on the results of this approach by itself. The best graduation for the weighted mortality rates was the model for which the CART/CIT tree was combined with the empirical approach.

Table 4.2: Test RMSE per Model (weighted mortality rates)

Model	**Theory**	**Results**	**Test RMSE**	**Rank**
Gompertz's law	2.2.1	4.1	0,0478	5
empirical approach	2.2.2	4.2	0,0463	2
standard table (GKM/F80)	2.2.3	4.3	0,0471	3
standard table (GKM/F95)	2.2.3	4.3	0,0473	4
GLM	2.2.4	4.4	0,3964	10
CART/CIT	2.2.5.1, 2.2.5.2	4.5.1, 4.5.2	0,0482	6
RF	2.2.5.3	4.5.3	0,0483	7
CART/CIT+ Gomp.	2.2.6	4.6.1, 4.6.4	0,0494	8
CART/CIT+ emp.	2.2.6	4.6.2, 4.6.5	0,0461	1
CART/CIT + GLM	2.2.6	4.6.3, 4.6.6	0,3464	9

Chapter 5

Actuarial Application

In this chapter, we will develop an actuarial application of the best models according to the previous section: CART combined with the empirical approach (Subsection 4.6.2) for the traditional mortality rates and CART/CIT combined with the empirical approach (Subsections 4.6.2, 4.6.5) for weighted mortality rates. More specifically, we will estimate claim costs for a one year policy for one insured life with 100.000€ of sum assured.

The best estimate for the claim amount of a one year life insurance for a given age x, considering a mortality rate q_x, an interest rate of 2% and 100.000€ of sum assured, where $v = \frac{1}{1+i} = \frac{1}{1+2\%} = 0,9804$, is (see Bowers *et al.* 1997):

$$BE = 100.000 \times q_x \times v^{1/2}. \tag{5.0.1}$$

To analyse the behaviour of the models, we will break the examples into four different ages and within each vary the remaining characteristics (civil status, NUTS_M, gender). These examples are meant to briefly illustrate the impact of the explanatory variables in each model. For the purpose of this exercise, consider the re-arranging of the leaf numbers in Table 5.1 (m.r. is short for mortality rates).

Table 5.1: Partitions for the trees in the best models

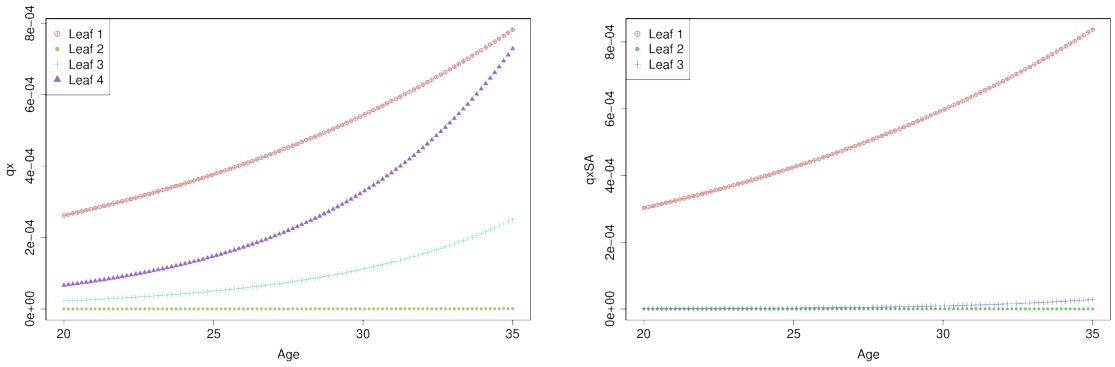
Partition	**Leaf (traditional m.r.)**	**Leaf (weighted m.r.)**
$NUTS\_M \neq Norte$	1	1
$NUTS\_M = Norte \wedge civil\ status \neq single$	2	2
$NUTS\_M = Norte \wedge civil\ status = single \wedge gender = male$	3	3
$NUTS\_M = Norte \wedge civil\ status = single \wedge gender = female$	4	

The values of the best estimates presented bellow will have differences from the values calculated if the exact equations in the Chapter 4 were used because of rounding. For the values bellow, the full capability of R regarding significant digits was used, leading to more accurate predictions.

Ages under 35

Figure 5.1 shows the best models broken down by leaf and for ages under 35.

§



(a) (b)

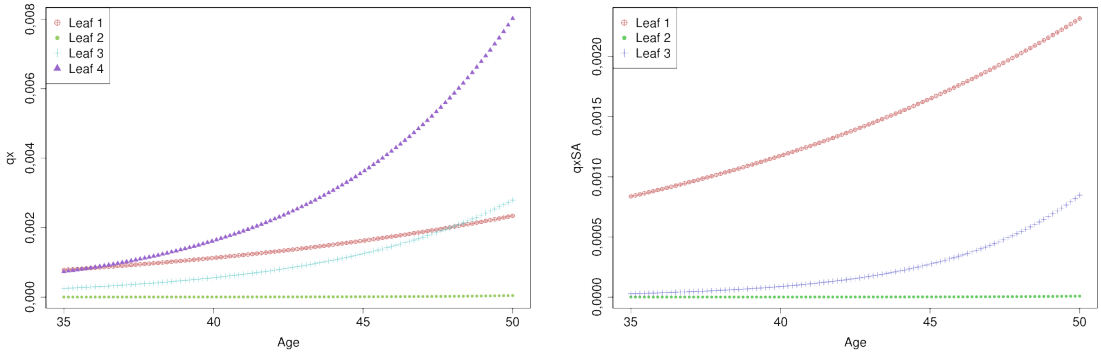
Figure 5.1: Fitted curves for ages under 35 for m.r.: (a) traditional; (b) weighted.

If the insured person is a 34 year old male widow from Algarve (leaf 1 in both models) using equation 5.0.1, the best estimate for the claim amount is 72€ considering the traditional m.r. and 77€ considering the weighted version. If the same person is from NUTS_M Norte, belonging to leaf 2 in both models, then BE is 0,042€ considering the traditional m.r. and 0,003€ considering the weighted version. Looking at the plots in Figure 5.1, we can see that the rates for leaf 2 are significantly lower than for leaf 1. Considering now that the same insured life is single instead of widowed (leaf 3 for both models), the best estimate goes back up to 21€ for the traditional m.r. and 2€ for the weighted m.r.. Lastly, if that person is female instead of male, then she will belong to leaf 4 for the traditional m.r. model but remain in leaf 3 for the weighted version. The BE still is 2€ for the weighted m.r. and rises to 61€ for the traditional rates.

Ages 35 to 50

The plot of the best models broken down by leaf and for ages 35 to 50, for both traditional and weighted m.r., can be seen in Figure 5.2.

The BE of a 45 year old male widow from Algarve (leaf 1 in both models) is 161€ considering the traditional m.r. and 163€ considering the weighted version. The same person but with NUTS_M Norte (leaf 2 in both models) has a BE of 1€ for the traditional rates and 0,148€ for the weighted version. Considering that life is now single (leaf 3), the best estimate for the claim amount is 124€ for the traditional m.r. and 27€ for the weighted version. Finally, if it is a 45 year old female single from Norte (leaf 4 for the traditional rates and 3 for the weighted rates), we will have a BE of 357€ for the traditional rates and still 27€ for the weighted rates.

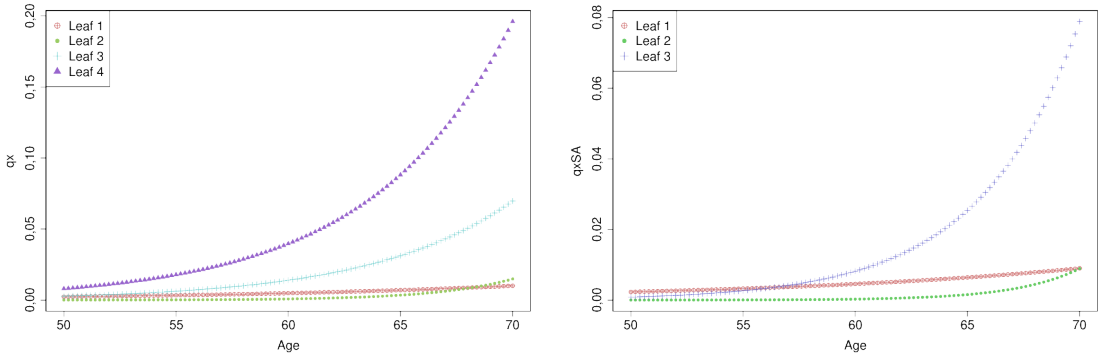


(a) (b)

Figure 5.2: Fitted curves for ages 35 to 50 for m.r.: (a) traditional; (b) weighted.

Ages 50 to 70

Figure 5.3 shows the plot of the best models broken down by leaf and for ages 50 to 70.



(a) (b)

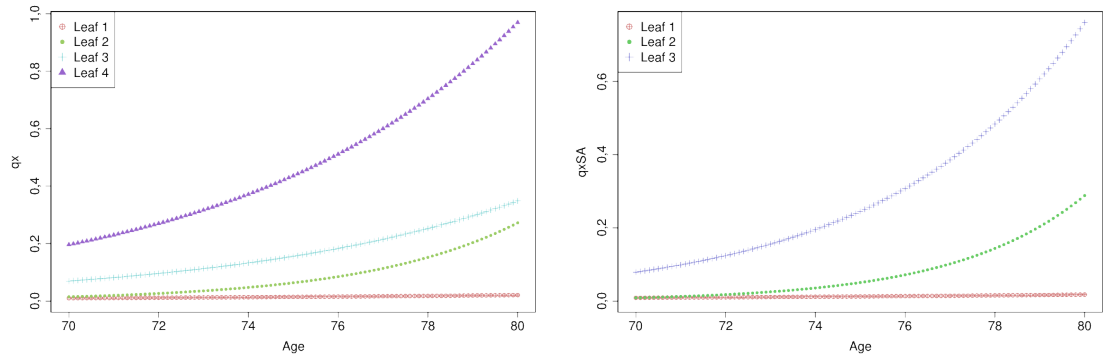
Figure 5.3: Fitted curves for ages 50 to 70 for m.r.: (a) traditional; (b) weighted.

Continuing with the analysis, as an example of leaf 1, we will now consider a 60 year old male widow from Algarve. The BE is then 480€ for the traditional m.r. and 452€ for the weighted version. If he is from Norte instead (leaf 2), then the BE is 80€ for the traditional m.r. and 27€ for the weighted ones. For leaf 3, the insured life would have to be single, making the best estimate 1.382€ for the traditional rates and 810€ for the weighted rates. Lastly, if the person is a female (leaf 4 for the traditional rates), nothing changes for the weighted mortality BE but the traditional version becomes 3.925€.

Ages over 70

Figure 5.4 shows, for both traditional and weighted m.r., the plot of the best models broken down by leaf and for ages over 70.

Consider a 75 year old male widowed individual from Algarve (leaf 1). Then, the best estimate for the claim amount is 1.436€ for the traditional m.r. and 1.251€ for the



(a) (b)

Figure 5.4: Fitted curves for ages over 70 for m.r.: (a) traditional; (b) weighted.

weighted version. If the same person is from NUTS_M Norte, belonging to leaf 2 in both models, then the BE is 6.298€ considering the traditional m.r. and 5.015€ considering the weighted version. Considering now that the same insured life is single(leaf 3 for both models), the best estimate goes up to 15.438€ for the traditional m.r. and 24.255€ for the weighted m.r.. Lastly, if that person is female, then she will belong to leaf 4 for the traditional m.r. models but remains in leaf 3 for the weighted version. Hence, the BE still is 24.255€ for the weighted m.r. and rises to 43.161€ for the traditional rates.

This chapter helps to illustrate how different the mortality behaviour is within each leaf, showing the potential for improvement in mortality estimation and ratemaking using the growing number of characteristics available to the insurer about its portfolio. More personalised tariffs could even be seen as fairer, given that each person would pay a price more in sync with their peers' mortality experience. However, this differentiation would also be very hard to justify to the person paying more and, in a way, the whole purpose of insurance will be lost.

The high differences in the BE between leafs is explained in part by the fact that leaf 1 encompasses 71% of the exposure for the traditional m.r. and 75% for the weighted version. The smaller size of the other leafs makes them more vulnerable to outliers' impact, such as the policy identified in Subsection 4.3.

These examples also provide insight into the impact of the sum assured in the estimates as the values from the two mortality approaches in this study (traditional and weighted by sum assured) are compared. While before it was stated (Subsection 4.6.2) that for these models the weighted m.r. were on average 88% of the traditional ones, in this chapter it can be observed that it depends on the leaf and ages.

For leaf 1, $\hat{q}_x^{SA}$ is higher than $\hat{q}_x$ until the age of 48 and is on average 99% of $\hat{q}_x$. For leaf 2, that average drops down to 30% ($\hat{q}_x^{SA}$ is higher than $\hat{q}_x$ only for ages 79 and 80) and goes back up to 55% for leaf 3 ($\hat{q}_x^{SA}$ is higher than $\hat{q}_x$ for ages over 68). Leaf 4 from the traditional rates' model is comparable to leaf 3 from the weighted rates' model and $\hat{q}_x^{SA}$ is on average 20% of $\hat{q}_x$, with no ages having higher $\hat{q}_x^{SA}$ than $\hat{q}_x$.

Chapter 6

Conclusions

The objective of this thesis was to study the problem of modelling mortality rates of a life insurer's portfolio. Several models were tested, each with regard to two perspectives: traditional mortality rates and mortality rates weighted by sum assured.

Before applying the different models, a preliminary analysis of the mortality rates by the different characteristics available (age, civil status, NUTS_M, gender) was performed in Section 3.4. The main conclusion of this analysis was that the mortality rates weighted by sum assured were, for the most part, lower than the corresponding traditional ones. This came as no surprise as lives with higher sums assured will likely be wealthier and have better health care, leading to later deaths.

In Chapter 4, some traditional methods were implemented (Gompertz's law, graduation by standard table, an empirical approach), followed by a generalised linear model. The use of regression trees (more specifically, CART, conditional inference trees and random forests) was introduced. Regression trees recursively sub-divide the space into smaller regions, where the interactions between explanatory variables are more manageable, and then fit simple models to each sub-space. Since the implemented regression trees assign a constant mortality rate to each leaf, hybrid methods were created, using the trees combined with some of the earlier models.

Several conclusions regarding the efficacy of the different models were true for both the traditional and weighted rates. Firstly, GLM (applied alone and with the trees) proved to be by far the worse method. While the models with GLM had an RMSE of over 30%, the remaining models had RMSE's of under 5%.

The empirical approach of simply fitting the mortality rates to an exponential curve, although not supported by the literature, proved to be a very efficient method alone and even better when used with CART.

The application of a simple transformation to standard tables, in this case a percentage, also presented very good results with the 80 series of the Swiss GK tables, proving to be a better fit than the 95 series. The results of this model in particular brought to light the impact of one specific policy which, due to its outlier behaviour, could had been excluded.

Regarding the remaining methods, the conclusions for the traditional and weighted

mortality rates differed. When modelling the traditional rates, random forest was the best of the tree generating algorithms, followed by the conditional inference trees and CART. Still better than CART applied alone, Gompertz's law proved to be on the lower end of the studied models in terms of accuracy. More surprising, applying Gompertz's law to the leafs of the CART and conditional inference trees was worse than fitting it for the whole training set at once.

For the modelling of the mortality rates weighted by sum assured, Gompertz's law was better than any of the tree models alone. Also, as explained before, CART and conditional inference trees generated the same results, which were better than random forest. Again, using Gompertz's law with the trees worsened its results.

For both approaches, the model considered best (lowest RMSE) was a tree with the empirical approach applied to its leafs. Looking at an example of the application of these two best models (Chapter 5), it was observed that the results vary immensely by considering the different combinations of the explanatory variables available.

Regarding possible improvements on the work, several suggestions can be made. To begin with, other hybrid models could be tried. One such example is the MOB algorithm (Zeileis *et al.* 2008), which also has a implementation in R by the same authors and focuses on model-based recursive partitioning.

The study could also be improved by using the available data for claim notification dates and taking into account the estimation of incurred but not reported (IBNR) claims. This information was not taken into consideration due to time constraints.

Other possible explanatory variables that could be tried are seniority in the portfolio (which would require taking into account possible contracts that started before 2011) and classes of sum assured (i.e., sums assured bellow 10.000€; sums assured between 10.000€ and 30.000€, etc). This second suggestion would show how much lower (or higher) the mortality rates are as the sum assured rises, which could be helpful in inferring discounts (or loadings) for people with higher sums assured.

Furthermore, as evidenced in the analysis of the results of graduation by standard table, there were some very high observed values of rates over the age of 70. So, another improvement could either be limiting the study at this age (instead of 80) or breaking down the models into two, for instance bellow and above 70 years old. This would prevent these high mortality values from affecting the models at younger ages.

Another interesting approach would be modelling the mortality rates in the context of the class imbalance problem, which is when the event to predict is a minority in the data. In our case, the ratio of claims over number of policies was roughly 5 thousand to 1,2 million. Extensive research has been done to mitigate this issue, with examples in Liu *et al.* (2010); Chawla (2005) and Chen and Breiman (2004). Techniques based in specialised supervised learning methods (bias in the learning algorithm in favour of the minority class) or biased sampling (training set where the frequency of both classes is the same) could be used.

Overall, this work shows the potential there is to create more personalised tariffs as one gets more information about the insured lives. Furthermore, even if some characteristics can't be used for directly for pricing due to legal constrains (in Portugal, for instance, insurance companies can't differentiate prices for man and woman), this sort of more detailed mortality tables can and should be used for product design and profit testing. Periodic repetition of a mortality study such as this will also give insight into the evolution of the mortality of the portfolio, which is valuable information for the insurer.

Bibliography

Bowers, N., Gerber, H., Hickman, J., Jones, D., Nesbitt, C., 1997. Actuarial Mathematics, 2nd Edition. The Society of Actuaries, Illinois.

Breiman, L., 2001. Random forests. In: Machine Learning.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Chapados, N., 2010. Data mining algorithms for actuarial ratemaking. Technological White Paper.

Chawla, N. V., 2005. Data Mining and Knowledge Discovery Handbook. Ch. Data Mining for Imbalanced Datasets: An Overview.

Chen, C., Breiman, L., 2004. Using Random Forest to Learn Imbalanced Data. Tech. rep.

Cody, D. C., 1941. Actuarial note: the standard deviation in the rate of mortality by amounts. Transactions, Actuarial Society of America (42), 69–73.

Debón, A., Montes, F., Sala, R., 2005. A comparison of parametric models for mortality graduation. application to mortality data for the valencia region (spain). SORT 29 (2), 269–288.

Gavin, J., Haberman, S., Verrall, R., 1993. Moving weighted average graduation using kernel estimation. Insurance: Mathematics and Economics 12 (2), 113–126.

Gompertz, B., 1825. On the Nature of the Function Expressive of the Law of Human Mortality: And on a New Mode of Determining the Value of Life Contingencies. In a Letter to Francis Baily. W. Nicol.

Guo, L., , Wang, M. C., 2002. Data mining techniques for mortality at advanced age. Tech. rep.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer New York Inc.

Haycocks, H., Perks, W., 1955. Mortality and other investigations. No. vol. 1. Published for the Institute of Actuaries and the Faculty of Actuaries at the University Press.

Hothorn, T., Hornik, K., Zeileis, A., 2014. ctree: Conditional inference trees. R News.

Hothorn, T., Hornik, K., Zeileis, A., Wien, W., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Statist.

Instituto Nacional de Estatística, 2013. As novas unidades territoriais para fins estatísticos. Brochure.

Kim, S., Kim, W., Park, R., 2011. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Research.

Klugman, S. A., 1981. On the variance and mean squared error of decrement estimators. Transactions, Actuarial Society of America (33), 301–31.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News.

Liu, W., Chawla, S., Cieslak, D. A., Chawla, N. V., 2010. A robust decision tree algorithm for imbalanced data sets. In: in SIAM International Conference on Data Mining, 2010. pp. 766–777.

Makeham, W., 1860. On the law of mortality and the construction of annuity tables. Journal of the Institute of Actuaries 8, 301–310.

McCullagh, P., Nelder, J. A., 1989. Generalized linear models (Second edition). London: Chapman & Hall.

Murthy, S. K., 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. Data Min. Knowl. Discov., 345–389.

Oppermann, L., 1872. Insurance record 1870. Journal of the Institute of Actuaries 16, 335–353.

Renshaw, A. E., 1991. Actuarial graduation practice and generalised linear models. Journal of the Institute of Actuaries (1886-1994), 295–312.

Roberts, L. A., 1992. On ratios of random variables and generalised mortality rates. Journal of Applied Probability, 268–279.

Roberts, L. A., 1993. Weighted mortality rates as early warning signals for insurance. ASTIN Bulletin, Workshop, Institute of Statistics and Operations Research, Victoria University.

Safavian, S. R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics 21 (3).

SAPS Mortality Committee, 2008. Methodology and assumptions used for cmi self-administered pension schemes mortality experience analyses. Working Paper 34.

SAPS Mortality Committee, 2009. The graduations of the cmi self-administered pension schemes 2000-2006 mortality experience. Working Paper 35.

Tan, P.-N., Steinbach, M., Kumar, V., 2006. Introduction to data mining. Pearson Addison Wesley.

Therneau, T. M., Atkinson, E. J., Foundation, M., 2015. An introduction to recursive partitioning using the rpart routines. R News.

Thiele, T. N., 1871. On a mathematical formula to express the rate of mortality throughout the whole of life. Journal of the Institute of Actuaries and Assurance Magazine 16, 313–329.

Verrall, R. J., 1996. A unified framework for graduation. Working Paper Actuarial Research Paper No. 91, City University London, London, UK.

Zeileis, A., Hothorn, T., Hornik, K., 2008. Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17.

Appendix A

Standard tables

Age	GKM80	GKF80	GKM95	GKF95
20	0,00114	0,0003262	0,0015503	0,001055
21	0,001152	0,0003364	0,0015094	0,001067
22	0,001164	0,0003613	0,0014643	0,001079
23	0,001177	0,000391	0,0014238	0,001091
24	0,001189	0,0004213	0,001388	0,001103
25	0,001201	0,0004514	0,0013574	0,001116
26	0,001213	0,0004819	0,0013325	0,001128
27	0,001225	0,0005129	0,0013137	0,00114
28	0,001238	0,0005448	0,0013018	0,001152
29	0,00125	0,000578	0,0012968	0,001164
30	0,001262	0,0006126	0,0012995	0,001177
31	0,001276	0,0006492	0,0013104	0,001189
32	0,001299	0,0006877	0,0013299	0,001201
33	0,00134	0,0007287	0,0013586	0,001213
34	0,001398	0,0007726	0,001397	0,001225
35	0,001477	0,0008193	0,0014454	0,001238
36	0,001577	0,0008693	0,0015045	0,00125
37	0,0017	0,0009216	0,0015754	0,001262
38	0,001847	0,0009756	0,0016591	0,001276
39	0,002021	0,0010304	0,0017566	0,001299
40	0,002222	0,001085	0,0018694	0,00134
41	0,002452	0,0011389	0,0019983	0,001398
42	0,002712	0,0011911	0,0021445	0,001477
43	0,003004	0,0012416	0,0023096	0,001577
44	0,00333	0,0012937	0,002497	0,0017
45	0,003691	0,0013517	0,0027107	0,001847
46	0,004089	0,0014197	0,0029545	0,002021
47	0,004524	0,001502	0,0032325	0,002222
48	0,005	0,0016022	0,0035482	0,002452
49	0,005516	0,0017249	0,0039057	0,002712
50	0,006094	0,0018738	0,0043087	0,003005

Age	GKM80	GKF80	GKM95	GKF95
51	0,006706	0,0020531	0,0047606	0,003331
52	0,007382	0,0022649	0,0052655	0,003691
53	0,00813	0,0025056	0,0058269	0,004089
54	0,008958	0,0027701	0,0064474	0,004525
55	0,009872	0,0030534	0,0071294	0,005
56	0,010882	0,0033504	0,0078756	0,005516
57	0,011998	0,003656	0,0086884	0,006094
58	0,013231	0,0039654	0,0095704	0,006706
59	0,014591	0,0042734	0,0105241	0,007382
60	0,016093	0,0045752	0,0115521	0,00813
61	0,017749	0,0048654	0,0126571	0,008958
62	0,019575	0,0051379	0,0138417	0,009872
63	0,021587	0,0055084	0,0151083	0,010882
64	0,023803	0,00609	0,0164598	0,011998
65	0,026242	0,0068875	0,0180706	0,013231
66	0,028925	0,0079057	0,0200313	0,014591
67	0,031873	0,0091493	0,0223416	0,016093
68	0,035111	0,0106232	0,0250018	0,017749
69	0,038662	0,012332	0,0280117	0,019575
70	0,042555	0,0142806	0,0313714	0,021587
71	0,046816	0,0164736	0,0350808	0,023803
72	0,051476	0,018916	0,03914	0,026242
73	0,056564	0,0216123	0,043549	0,028925
74	0,062113	0,0245675	0,0483078	0,031873
75	0,068153	0,0277862	0,0534163	0,035111
76	0,074718	0,0312732	0,0588745	0,038662
77	0,081839	0,0350334	0,0646826	0,042555
78	0,089548	0,0390713	0,0708404	0,046817
79	0,097876	0,0433919	0,077348	0,051476
80	0,10685	0,0479999	0,0842053	0,056565

Appendix B

Database transformation algorithm

1. Define DB_dates as the data set containing all unique combinations of (PersonKey, DateBirth, PolicyDates), where PolicyDates are the DateIssue and EndDate for the policy in question.
2. Define DB_persons as the dataset containing (PersonKey, Min(PolicyDates), Max(PolicyDates)), where Min is the minimum and Max is the maximum of PolicyDates, grouped by PersonKey.
3. Define an empty data set named Intervals;
4. For each PersonKey in DB_persons:
   (a) Define ListDates as the subset of DB_dates for the PersonKey in question;
   (b) Add to ListDates the person's birthdays contained inside the (Min(PolicyDates), Max(PolicyDates)) interval (this will allow tracking age changes);
   (c) Add to ListDates the $01/01/Year$, where $Year \in \{2011, 2012, 2013, 2014\}$, which are inside the (Min(PolicyDates), Max(PolicyDates)) interval (this will allow tracking sum assured changes);
   (d) Re-define ListDates as the sorted version of the unique values of the previously defined ListDates;
   (e) Define IntervalsPers as the data set containing (PersonKey, IntervalStart, IntervalEnd), where:
      i. IntervalStart has the first until the one before the last member of ListDates;
      ii. IntervalEnd has the second until last member of ListDates;
   (f) Append IntervalsPers to Intervals.
5. Join Intervals with the original data base (by PersonKey), adding only the lines for which (IntervalStart, IntervalEnd) are inside the (DateIssue, EndDate) interval of the policy in question and call this data base DB_new. The original data set will now have each policy repeated each time there is a change in the insured person's age or sum assured or there is an intersection with the timeframe of another policy

(with the same insured person).

6. For each line of BD_new:
   (a) Insert the variable SA (sum assured) defined as either SA_2011, SA_2012, SA_2013 or SA_2014, depending on the year of the IntervalStart variable of that line;
   (b) Calculate the risk exposure (named Ex) of that interval as the exact number of days between IntervalStart and IntervalEnd divided by the number of days in the year in question $\left(E_x = \frac{number\ days\ exposed\ risk\ (age\ x)}{number\ days\ year\ (IntervalStart)}\right)$;
   (c) Calculate ExSA as variable Ex times variable SA $\left(E_x^{SA} = E_x \times SA\right)$;
   (d) Calculate the real age at the start of the interval (variable Age);
   (e) Define the variable Claim as:
      i. If EndDate (of the policy) is equal to the IntervalEnd in question and the policy had a death claim, Claim= 1.
      ii. Otherwise, Claim= 0;
   (f) Define the variable SA_Claim as the product of SA and Claim ($SA\_Claim = SA \times Claim$), representing the sum assured at the time death occurred.

Appendix C

Stratification Algorithm

1. For the training set:
   (a) Calculate the number of unique insured lives in each distinct combination of (gender, civil status, NUTS_M) and multiply it by 0,8, generating a vector containing the number of people to be randomly selected from each stratum (gender, civil status, NUTS_M);
   (b) Join each stratum (gender, civil status, NUTS_M) with a data set containing each unique person and their characteristics and apply random sampling to that subset, without replacement, until the required number of samples is reached.
2. For the test set, utilise the remaining, unpicked population.